COSMOS QA: Machine Reading Comprehension with Contextual Commonsense Reasoning

Lifu Huang[♠*], **Ronan Le Bras**[♣], **Chandra Bhagavatula**[♣], **Yejin Choi**[♣♢]
[♠] University of Illinois Urbana-Champaign, Champaign, IL, USA
[♣] Allen Institute for Artificial Intelligence, Seattle, WA, USA
[♢] University of Washington, Seattle, WA, USA
lifuh2@illinois.edu
{ronanlb, chandrab, yejinc}@allenai.org

Abstract

Understanding narratives requires reading between the lines, which in turn, requires interpreting the likely causes and effects of events, even when they are not mentioned explicitly. In this paper, we introduce COSMOS QA, a large-scale dataset of 35,600 problems that require commonsense-based reading comprehension, formulated as multiple-choice questions. In stark contrast to most existing reading comprehension datasets where the questions focus on factual and literal understanding of the context paragraph, our dataset focuses on reading between the lines over a diverse collection of people's everyday narratives, asking such questions as *"what might be the possible reason of ...?"*, or *"what would have happened if ..."* that require reasoning beyond the exact text spans in the context. To establish baseline performances on COSMOS QA, we experiment with several state-of-the-art neural architectures for reading comprehension, and also propose a new architecture that improves over the competitive baselines. Experimental results demonstrate a significant gap between machine (68.4%) and human performance (94%), pointing to avenues for future research on commonsense machine comprehension. Dataset, code and leaderboard is publicly available at https://wilburone.github.io/cosmos.

P1: It's a very humbling experience when you need someone to dress you every morning, tie your shoes, and put your hair up. Every menial task takes an unprecedented amount of effort. It made me appreciate Dan even more. But anyway I shan't dwell on this (I'm not dying after all) and not let it detract from my lovely 5 days with my friends visiting from Jersey.

Q: *What's a possible reason the writer needed someone to dress him every morning?*

A: The writer doesn't like putting effort into these tasks.
✓ **B: The writer has a physical disability.**
C: The writer is bad at doing his own hair.
D: None of the above choices.

P2: A woman had topped herself by jumping off the roof of the hospital she had just recently been admitted to. She was there because the first or perhaps latest suicide attempt was unsuccessful. She put her clothes on, folded the hospital gown and made the bed. She walked through the unit unimpeded and took the elevator to the top floor.

Q: *What would have happened to the woman if the staff at the hospital were doing their job properly?*

✓ **A: The woman would have been stopped before she left to take the elevator to the top floor and she would have lived.**
B: She would have been ushered to the elevator with some company.
C: She would have managed to get to the elevator quicker with some assistance.
D: None of the above choices.

Figure 1: Examples of COSMOS QA. (✓ indicates the correct answer.) Importantly, (1) the correct answer is not explicitly mentioned anywhere in the context paragraph, thus requiring reading between the lines through commonsense inference and (2) answering the question correctly requires reading the context paragraph, thus requiring reading comprehension and *contextual* commonsense reasoning.

1 Introduction

Reading comprehension requires not only understanding what is stated explicitly in text, but also *reading between the lines*, i.e., understanding what is not stated yet obviously true (Norvig, 1987).

For example, after reading the first paragraph in Figure 1, we can understand that the writer is not a child, yet needs someone to dress him or her every morning, and appears frustrated with the current situation. Combining these clues, we can infer that the plausible reason for the writer being dressed by other people is that he or she may have a physical disability.

As another example, in the second paragraph of Figure 1, we can infer that the woman was admitted to a psychiatric hospital, although not mentioned explicitly in text, and also that the job of the hospital staff is to stop patients from committing suicide. Furthermore, what the staff should

[*]The work has been done during the author's internship in AI2.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 2391–2401, Hong Kong, China, November 3–7, 2019.

have done, in the specific situation described, was to stop the woman from taking the elevator.

There are two important characteristics of the problems presented in Figure 1. First, the correct answers are not explicitly mentioned anywhere in the context paragraphs, thus requiring reading between the lines through commonsense inference. Second, selecting the correct answer requires reading the context paragraphs. That is, if we were not provided with the context paragraph for the second problem, for example, the plausible correct answer could have been B or C instead.

In this paper, we focus on reading comprehension that requires *contextual* commonsense reasoning, as illustrated in the examples in Figure 1. Such reading comprehension is an important aspect of how people read and comprehend text, and yet, relatively less studied in the prior machine reading literature. To support research toward commonsense reading comprehension, we introduce COSMOS QA (**Co**mmon**s**ense **M**achine C**o**mprehen**s**ion), a new dataset with $35,588$ reading comprehension problems that require reasoning about the causes and effects of events, the likely facts about people and objects in the scene, and hypotheticals and counterfactuals. Our dataset covers a diverse range of everyday situations, with $21,886$ distinct contexts taken from blogs of personal narratives.

The vast majority (93.8%) of our dataset requires contextual commonsense reasoning, in contrast with existing machine comprehension (MRC) datasets such as SQuAD (Rajpurkar et al., 2016), RACE (Lai et al., 2017), Narrative QA (Kočiskỳ et al., 2018), and MCScript (Ostermann et al., 2018), where only a relatively smaller portion of the questions (e.g., 27.4% in MCScript) require commonsense inference. In addition, the correct answer cannot be found in the context paragraph as a text span, thus we formulate the task as multiple-choice questions for easy and robust evaluation. However, our dataset can also be used for generative evaluation, as will be demonstrated in our empirical study.

To establish baseline performances on COSMOS QA, we explore several state-of-the-art neural models developed for reading comprehension. Furthermore, we propose a new architecture variant that is better suited for commonsense-driven reading comprehension. Still, experimental results demonstrate a significant gap between machine (68.4% accuracy) and human performance (94.0%). We provide detailed analysis to provide insights into potentially promising research directions.

2 Dataset Design

2.1 Context Paragraphs

We gather a diverse collection of everyday situations from a corpus of personal narratives (Gordon and Swanson, 2009) from the Spinn3r Blog Dataset (Burton et al., 2009). Appendix A provides additional details on data pre-processing.

2.2 Question and Answer Collection

We use Amazon Mechanical Turk (AMT) to collect questions and answers. Specifically, for each paragraph, we ask a worker to craft at most two questions that are related to the context and require commonsense knowledge. We encourage the workers to craft questions from but not limited to the following four categories:

Causes of events: What may (or may not) be the plausible reason for an event?

Effects of events: What may (or may not) happen before (or after, or during) an event?

Facts about entities: What may (or may not) be a plausible fact about someone or something?

Counterfactuals: What may (or may not) happen if an event happens (or did not happen)?

These 4 categories of questions literally cover all 9 types of social commonsense of Sap et al. (2018). Moreover, the resulting commonsense also aligns with 19 ConceptNet relations, e.g., *Causes*, *HasPrerequisite* and *MotivatedByGoal*, covering about 67.8% of ConceptNet types. For each question, we also ask a worker to craft at most two correct answers and three incorrect answers. We paid workers $0.7 per paragraph, which is about $14.8 per hour. Appendix B provides additional details on AMT instructions.

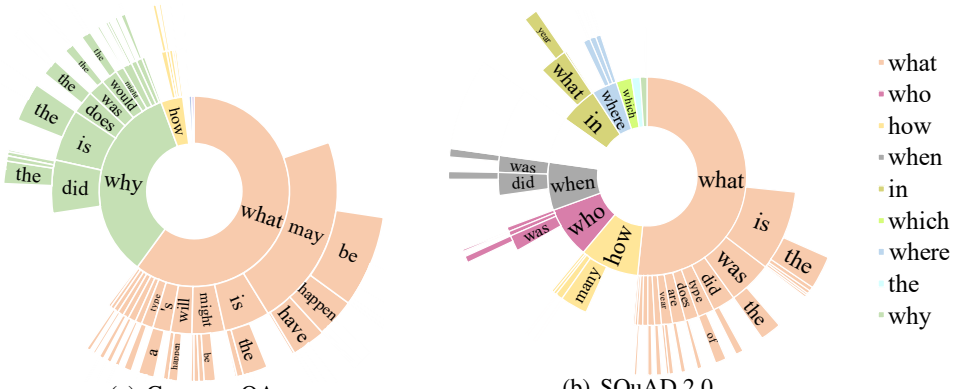
2.3 Validation

We create multiple tasks to have humans verify the data. Given a paragraph, a question, a correct answer and three incorrect answers,[1] we ask AMT workers to determine the following sequence of questions: (1) whether the paragraph is inappropriate or nonsensical, (2) whether the question is

[1] If a question is crafted with two correct answers, we will create two question sets with each correct answer and the same three incorrect answers.

	Train	Dev	Test	All
# Questions (Paragraphs)	25,588 (13,715)	3,000 (2,460)	7,000 (5,711)	35,588 (21,866)
Ave./Max. # Tokens / Paragraph	69.4 / 152	72.6 / 150	73.1 / 149	70.3/ 152
Ave./Max. # Tokens / Question	10.3 / 34	11.2 / 28	11.2 / 29	10.6 / 34
Ave./Max. # Tokens / Correct Answer	8.0 / 40	9.7 / 41	9.7 / 36	8.5 / 41
Ave./Max. # Tokens / Incorrect Answer	7.6 / 40	9.1 / 38	9.1 / 36	8.0 / 40
Percentage of Unanswerable Questions	5.9%	8.7%	8.4%	6.7%

Table 1: Statistics of training, dev and test sets of COSMOS QA.



(a) COSMOS QA (b) SQuAD 2.0

Figure 2: Distribution of trigram prefixes of questions in COSMOS QA and SQuAD 2.0

nonsensical or not related to the paragraph, (3) whether they can determine the most plausible correct answer, (4) if they can determine the correct answer, whether the answer requires commonsense knowledge, and (5) if they can determine the correct answer, whether the answer can be determined without looking at the paragraph.

We follow the same criterion as in Section 2.2 and ask 3 workers to work on each question set. Workers are paid $0.1 per question. We consider as valid question set where at least two workers correctly picked the intended answer and all of the workers determined the paragraph/question/answers as satisfactory. Finally we obtain 33,219 valid question sets in total.

2.4 Unanswerable Question Creation

With human validation, we also obtain a set of questions for which workers can easily determine the correct answer without looking at the context or using commonsense knowledge. To take advantage of such questions and encourage AI systems to be more consistent with human understanding, we create unanswerable questions for COSMOS QA. Specifically, from validation outputs, we collect 2,369 questions for which at least two workers correctly picked the answer and at least on worker determined that it is answerable without looking at the context or requires no common sense. We replace the correct choice of these questions with a "None of the above" choice.

To create false negative training instances, we randomly sample 70% of questions from the 33,219 good question sets and replace their least challenging negative answer with "None of the above". Specifically, we fine-tune three BERT[2] next sentence prediction models on COSMOS: BERT($A|P,Q$), BERT($A|P$), BERT($A|Q$), where P, Q, A denotes the paragraph, question, and answer. BERT($A|\triangle$) denotes the possibility of an answer A being the next sentence of $\triangle$. The least challenging negative answer is determined by

$$A^{'} = \arg\min\Big(\sum_{\forall \triangle \subseteq \{P,Q\}} \text{BERT}(A|\triangle)\Big)$$

2.5 Train / Dev / Test Split

We finally obtain 35,588 question sets for our COSMOS dataset. To ensure that the development and test sets are of high quality, we identify a group of workers who excelled in the generation task for question and answers, and randomly sample 7K question sets authored by these excellent workers as test set, and 3K question sets as development set. The remaining questions are all used as training set. Table 1 shows dataset statistics.

[2]Through the whole paper, BERT refers to the pre-trained BERT large uncased model from https://github.com/huggingface/pytorch-pretrained-BERT

2.6 Data Analysis

Figure 2 compares frequent trigram prefixes in COSMOS and SQuAD 2.0 (Rajpurkar et al., 2018). Most of the frequent trigram prefixes in COSMOS, e.g., *why*, *what may happen*, *what will happen* are almost absent from SQuAD 2.0, which demonstrates the unique challenge our dataset contributes. We randomly sample 500 answerable questions to manually categorize according to their contextual commonsense reasoning types. Figure 3 shows representative examples. Table 2 shows the distribution of the question types.

- **Pre-/post-conditions:** causes/effects of an event.
- **Motivations:** intents or purposes.
- **Reactions:** possible reactions of people or objects to an event.
- **Temporal events:** what events might happen before or after the current event.
- **Situational facts:** facts that can be inferred from the description of a particular situation.
- **Counterfactuals:** what might happen given a counterfactual condition.
- **Other:** other types, e.g., cultural norms.

Type	Percentage (%)
MRC w/o commonsense	6.2
MRC w/ commonsense	93.8
Pre-/Post- Condition	27.2
Motivation	16.0
Reaction	13.2
Temporal Events	12.4
Situational Fact	23.8
Counterfactual	4.4
Other	12.6

Table 2: The distribution of contextual commonsense reasoning types in COSMOS.

3 Model

3.1 BERT with Multiway Attention

Multiway attention (Wang et al., 2018a; Zhu et al., 2018) has been shown to be effective in capturing the interactions between each pair of input paragraph, question and candidate answers, leading to better context interpretation, while BERT fine-tuning (Devlin et al., 2018) also shows its prominent ability in commonsense inference. To further enhance the context understanding ability of BERT fine-tuning, we perform multiway bidirectional attention over the BERT encoding output. Figure 4 shows the overview of the architecture.

Encoding with Pre-trained BERT Given a paragraph, a question, and a set of candidate answers, the goal is to select the most plausible correct answer from the candidates. We formulate the input paragraph as $P = \{p_0, p_1, ..., p_n\}$, the question as $Q = \{q_0, q_1, ..., q_k\}$ and a candidate answer as $A = \{a_0, a_1, ..., a_s\}$, where p_i, q_i and a_i is the i-th word of the paragraph, question and candidate answer respectively. Following (Devlin et al., 2018), given the input P, Q and A, we apply the same tokenizer and concatenate all tokens as a new sequence $\big[\text{[CLS]}, P, \text{[SEP]}, Q, \text{[SEP]}, A, \text{[SEP]}\big]$, where [CLS] is a special token used for classification and [SEP] is a delimiter. Each token is initialized with a vector by summing the corresponding token, segment and position embedding from pre-trained BERT, and then encoded into a hidden state. Finally we get $[\mathbf{H}_{cls}, \mathbf{H}_P, \mathbf{H}_Q, \mathbf{H}_A]$ as encoding output.

Multiway Attention To encourage better context interpretation, we perform multiway attention over BERT encoding output. Taking the paragraph P as an example, we compute three types of attention weights to capture its correlation to the question, the answer, and both the question and answer, and get question-attentive, answer-attentive, and question and answer-attentive paragraph representations

$$\tilde{\mathbf{H}}_P = \mathbf{H}_P \boldsymbol{W}_t + \boldsymbol{b}_t$$
$$\mathbf{M}_P^Q = \text{Softmax}(\tilde{\mathbf{H}}_P \mathbf{H}_Q^\top)\mathbf{H}_Q$$
$$\mathbf{M}_P^A = \text{Softmax}(\tilde{\mathbf{H}}_P \mathbf{H}_A^\top)\mathbf{H}_A$$
$$\mathbf{M}_P^{QA} = \text{Softmax}(\tilde{\mathbf{H}}_P \mathbf{H}_{QA}^\top)\mathbf{H}_{QA}$$

where $\boldsymbol{W}_t$ and $\boldsymbol{b}_t$ are learnable parameters. Next we fuse these representations with the original encoding output of P

$$\mathbf{F}_P^Q = \sigma([\mathbf{H}_P \mathbf{M}_P^Q : \mathbf{H}_P - \mathbf{M}_P^Q]\boldsymbol{W}_P + \boldsymbol{b}_P)$$
$$\mathbf{F}_P^A = \sigma([\mathbf{H}_P \mathbf{M}_P^A : \mathbf{H}_P - \mathbf{M}_P^A]\boldsymbol{W}_P + \boldsymbol{b}_P)$$
$$\mathbf{F}_P^{QA} = \sigma([\mathbf{H}_P \mathbf{M}_P^{QA} : \mathbf{H}_P - \mathbf{M}_P^{QA}]\boldsymbol{W}_P + \boldsymbol{b}_P)$$

where $[:]$ denotes concatenation operation. $\boldsymbol{W}_P$, $\boldsymbol{b}_P$ are learnable parameters for fusing paragraph representations. σ denotes ReLU function.

Finally, we apply column-wise max pooling on $[\mathbf{F}_P^Q : \mathbf{F}_P^A : \mathbf{F}_P^{QA}]$ and obtain the new paragraph representation $\mathbf{F}_P$. Similarly, we can also obtain a new representation $\mathbf{F}_Q$ and $\mathbf{F}_A$ for Q and A respectively. We use $\mathbf{F} = [\mathbf{F}_P : \mathbf{F}_Q : \mathbf{F}_A]$ as the

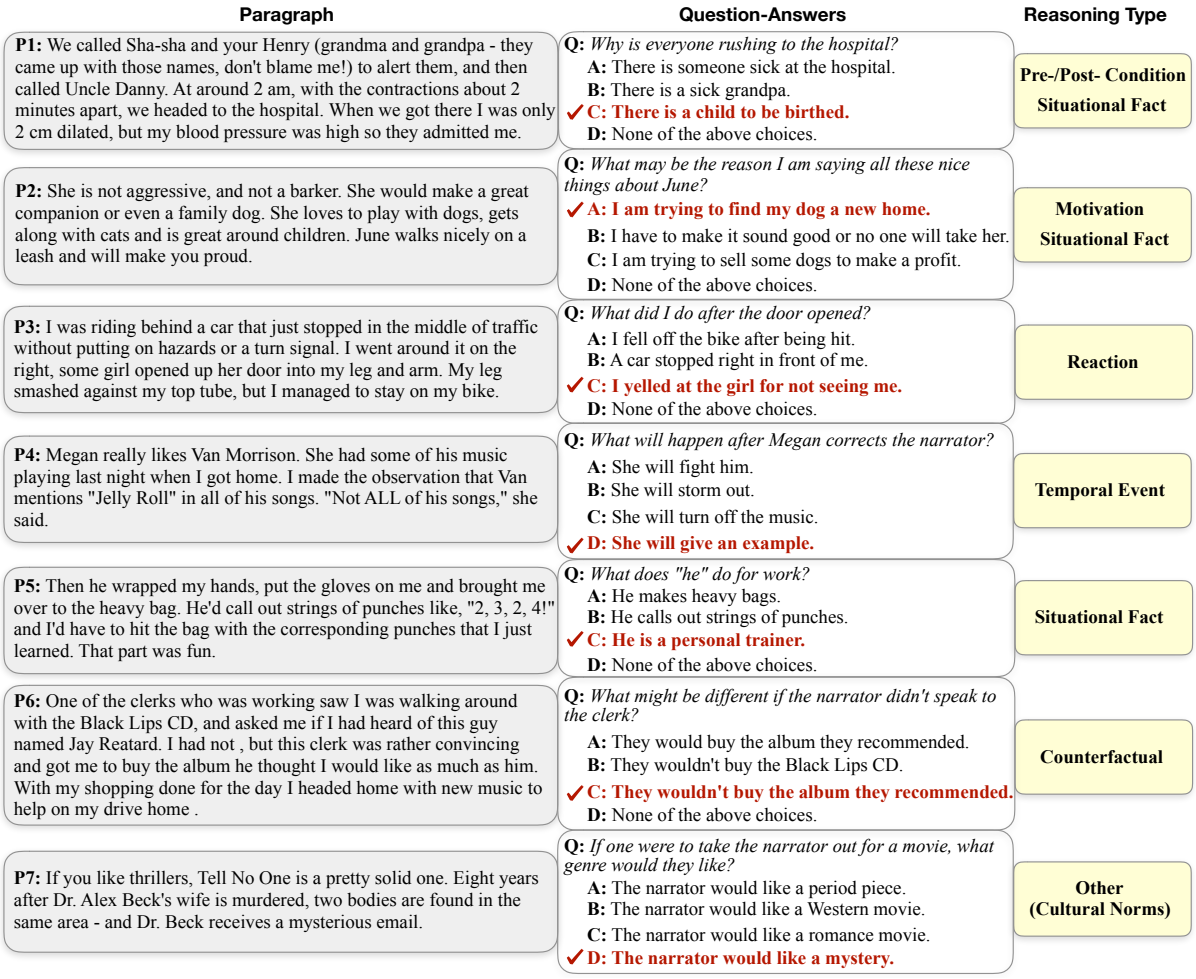


Figure 3: Examples of each type of commonsense reasoning in COSMOS QA. (✓ indicates the correct answer.)

overall vector representation for the set of paragraph, question and a particular candidate answer.

Classification For each candidate answer A_i, we compute the loss as follows:

$$L(A_i|P,Q) = -\log \frac{\exp(\boldsymbol{W}_f^\top \mathbf{F}_i)}{\sum_{j=1}^{4} \exp(\boldsymbol{W}_f^\top \mathbf{F}_j))}$$

4 Experiments

4.1 Baseline Methods

We explore two categories of baseline methods: reading comprehension approaches and pre-trained language model based approaches.

Sliding Window (Richardson et al., 2013) measures the similarity of each candidate answer with each window with m words of the paragraph.

Stanford Attentive Reader (Chen et al., 2016) performs a bilinear attention between the question and paragraph for answer prediction.

Gated-Attention Reader (Dhingra et al., 2017) performs multi-hop attention between the question and a recurrent neural network based paragraph encoding states.

Co-Matching (Wang et al., 2018b) captures the interactions between question and paragraph, as well as answer and paragraph with attention.

Commonsense-RC (Wang et al., 2018a) applies three-way unidirectional attention to model interactions between paragraph, question, and answers.

GPT-FT (Radford et al., 2018) is based on a generative pre-trained transformer language model, following a fine-tuning step on COSMOS QA.

BERT-FT (Devlin et al., 2018) is a pre-trained bidirectional transformer language model following a fine-tuning step on COSMOS QA.

DMCN (Zhang et al., 2019a) performs dual attention between paragraph and question/answer over BERT encoding output.

Human Performance To get human performance on COSMOS QA, we randomly sample 200 question sets from the test set, and ask 3 workers

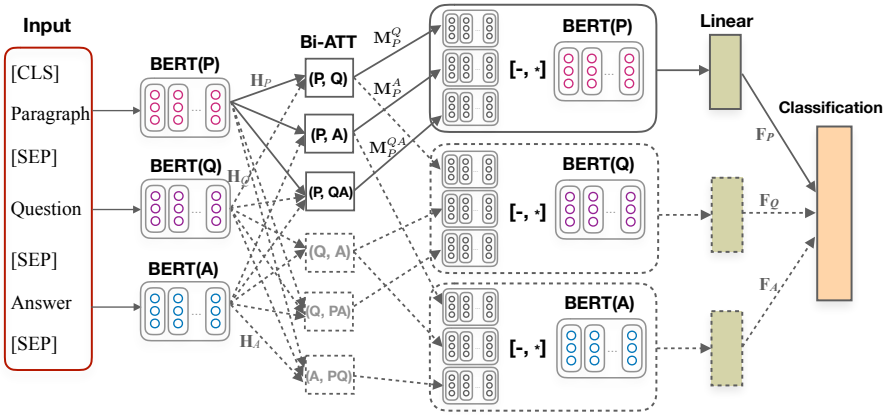


Figure 4: Architecture overview of BERT with multiway attention: Solid lines and blocks show the learning of multiway attentive context paragraph representation.

Model	Att(P, Q)	Att(P, A)	Att(Q, A)	Pre-training LM	Dev	Test
Sliding Window (Richardson et al., 2013)	✗	✗	✗	✗	25.0	24.9
Stanford Attentive Reader (Chen et al., 2016)	UD	✗	✗	✗	45.3	44.4
Gated-Attention Reader (Dhingra et al., 2017)	Multi-hop UD	✗	✗	✗	46.9	46.2
Co-Matching (Wang et al., 2018b)	UD	UD	✗	✗	45.9	44.7
Commonsense-Rc (Wang et al., 2018a)	UD	UD	UD	✗	47.6	48.2
GPT-FT (Radford et al., 2018)	✗	✗	✗	UD	54.0	54.4
BERT-FT (Devlin et al., 2018)	✗	✗	✗	BD	66.2	67.1
DMCN (Zhang et al., 2019a)	UD	UD	✗	BD	67.1	67.6
BERT-FT Multiway	BD	BD	BD	BD	68.3	68.4
Human						94.0

Table 3: Comparison of varying approachs (Accuracy %). Att: Attention, UD: Unidirectional, BD: Bidirectional

from AMT to select the most plausible correct answer. Each worker is paid $0.1 per question set. We finally combine the predictions for each question with majority vote.

4.2 Results and Analysis

Table 3 shows the characteristics and performance of varying approaches and human performance.[3]

Most of the reading comprehension approaches apply attention to capture the correlation between paragraph, question and each candidate answer and tend to select the answer which is the most semantically closed to the paragraph. For example, in Figure 5, the Commonsense-RC baseline mistakenly selected the choice which has the most overlapped words with the paragraph without any commonsense reasoning. However, our analysis shows that more than 83% of correct answers in COSMOS QA are not stated in the given paragraphs, thus simply comparing the semantic relatedness doesn't work well.

Pre-trained language models with fine-tuning achieve more than 20% improvement over reading comprehension approaches. By performing attention over BERT-FT, the performance is further improved, which demonstrates our assumption that incorporating interactive attentions can further enhance the context interpretation of BERT-FT. For example, in Figure 5, BERT-FT mistakenly selected choice A which can be possibly entailed by the paragraph. However, by performing multiway attention to further enhance the interactive comprehension of context, question and answer, our approach successfully selected the correct answer.

5 Discussion

5.1 Ablation Study

Model	Dev Acc (%)	Test Acc (%)
BERT-FT (A|P, Q)	66.2	67.1
BERT-FT (A|P)	63.5	64.5
BERT-FT (A|Q)	56.2	55.9
BERT-FT (A)	40.3	40.3

Table 4: Ablation of Paragraphs (P) or Questions (Q)

Many recent studies have suggested the importance of measuring the dataset bias by checking

[3] Appendix C shows the implementation details.

P: I cleaned the two large bottom cupboards and threw a ton of old stuff away. Dustin's parents like to drop off boxes of food like we're refugees or something. It's always appreciated, and some of it is edible. Most of what I threw away was from last year when Dustin's great-aunt was moving into her new apartment home (retirement center) and they cleaned out her pantry.

Q: *What is the most likely reason that I decided to clean the cupboards ?*

✗ **A:** I was getting tired of having food in the house.

✓ **B: We were getting more food and needed to create room.**

✗ **C:** Dustin and I split up and I need to get rid of his old stuff.

D: None of the above choices.

Figure 5: Prediction comparison between our approach (B) with Commonsense-RC (C) and BERT-FT (A).

the model performance based on partial information of the problem (Gururangan et al., 2018; Cai et al., 2017). Therefore, we report problem ablation study in Table 4 using BERT-FT as a simple but powerful straw man approach. Most notably, ablating questions does not cause significant performance drop. Further investigation indicates that this is because the high-level question types, e.g., *what happens next*, *what happened before*, are not diverse, so that it is often possible to make a reasonable guess on what the question may have been based on the context and the answer set. Ablating other components of the problems cause more significant drops in performance.

5.2 Knowledge Transfer Through Fine-tuning

Recent studies (Howard and Ruder, 2018; Min et al., 2017; Devlin et al., 2018) have shown the benefit of fine-tuning on similar tasks or datasets for knowledge transfer. Considering the unique challenge of COSMOS, we explore two related multiple-choice datasets for knowledge transfer: RACE (Lai et al., 2017), a large-scale reading comprehension dataset, and SWAG (Zellers et al., 2018), a large-scale commonsense inference dataset. Specifically, we first fine-tune BERT on RACE or SWAG or both, and directly test on COSMOS to show the impact of knowledge transfer. Furthermore, we sequentially fine-tune BERT on both RACE or SWAG and COSMOS. As Table 5 shows, with direct knowledge transfer, RACE provides significant benefit than SWAG since COSMOS requires more understanding of the interaction between paragraph, question and each candidate answer. With sequentially fine-tuning, SWAG provides better performance, which indicates that with fine-tuning on SWAG, BERT can obtain better commonsense inference ability, which is also beneficial to COSMOS.

Model	Dev Acc	Test Acc
BERT-FT$_{\text{SWAG}}$	28.9	28.5
BERT-FT$_{\text{RACE}}$	42.0	42.5
BERT-FT$_{\text{RACE+SWAG}}$	44.2	45.1
BERT-FT$_{\text{SWAG}\rightarrow\text{Cosmos}}$	67.8	68.9
BERT-FT$_{\text{RACE}\rightarrow\text{Cosmos}}$	67.4	68.2
BERT-FT$_{\text{RACE+SWAG}\rightarrow\text{Cosmos}}$	67.1	68.7

Table 5: Knowledge transfer through fine-tuning. (%)

P1: A woman had topped herself by jumping off the roof of the hospital she had just recently been admitted to. She was there because the first or perhaps latest suicide attempt was unsuccessful. She put her clothes on, folded the hospital gown and made the bed. She walked through the unit unimpeded and took the elevator to the top floor.

Q: *What would have happened to the woman if the staff at the hospital were doing their job properly?*

✓ **A: The woman would have been stopped before she left to take the elevator to the top floor and she would have lived.**

B: She would have been ushered to the elevator with some company.

✗ **C: She would have managed to get to the elevator quicker with some assistance.**

D: None of the above choices.

P2: Like me, she had no family or friends who could help with childcare. So like me, she found a daycare center that met her part-time needs. In sharp contrast to my job as a (gasp!) writer for the evil MSM, her nursing job was deemed by the other moms to be useful and worthwhile --in fact, worth putting her baby into daycare for "just a few hours, what harm could it do?"

Q: *What would happened if she could not find a daycare?*

✓ **A: She would try to find a babysitter.**

B: She would take the baby to work.

✗ **C: She would leave the baby alone at home.**

D: None of the above choices.

P3: My head hurts. I had so much fun at a chat with some scrap friends last Saturday night that I forgot to sleep. I ended up crawling into bed around 7AM.

Q: *What may have happened if she did n't chat to her scrap friends ?*

A: She would have done some scrap thing at home.

✓ **B: She would not have gotten up with a headache.**

✗ **C: She would have been lonely and stayed up all night.**

D: None of the above choices.

Figure 6: Examples errors of our approach. (✓ indicates correct answers and ✗ shows prediction errors.)

5.3 Error Analysis

We randomly select 100 errors made by our approach from the dev set, and identify 4 phenomena:

Complex Context Understanding: In 30% of the errors, the context requires complicated cross-sentence interpretation and reasoning. Taking P1 in Figure 6 as an example, to correctly predict the answer, we need to combine the context information that *the woman attempted to suicide before but failed*, *she made the bed since she determined to leave*, and *she took the elevator and headed to the roof*, and infer that *the woman was attempting to suicide again*.

Inconsistent with Human Common Sense: In

33% of the errors, the model mistakenly selected the choice which is not consistent with human common sense. For example, in P2 of Figure 6, both choice A and choice C could be potentially correct answers. However, from human common sense, *it's not safe to leave a baby alone at home*.

Multi-turn Commonsense Inference: 19% of the errors are due to multi-turn commonsense inference. For example, in P3 of Figure 6, the model needs to first determine the cause of *headache* is that *she chatted with friends and forgot to sleep* using common sense. Further, with counterfactual reasoning, if *she didn't chat to her friends*, then *she wouldn't have gotten up with a headache*.

Unanswerable Questions: 14% of the errors are from unanswerable questions. The model cannot handle "None of the above" properly since it cannot be directly entailed by the given paragraph or the question. Instead, the model needs to compare the potential of all the other candidate choices.

5.4 Generative Evaluation

In real world, humans are usually asked to perform contextual commonsense reasoning without being provided with any candidate answers. To test machine for human-level intelligence, we leverage a state-of-the-art natural language generator GPT2 (Radford et al., 2019) to automatically generate an answer by reading the given paragraph and question. Specifically, we fine-tune a pre-trained GPT2 language model on all the $[Paragraph, Question, Correct\ Answer]$ of COSMOS training set, then given each $[Paragraph, Question]$ from test set, we use GPT2-FT to generate a plausible answer. We automatically evaluate the generated answers against human authored correct answers with varying metrics in Table 6. We also create a AMT task to have 3 workers select all plausible answers among 4 automatically generated answers and a "None of the aboce" choice for 200 question sets. We consider an answer as correct only if all 3 workers determined it as correct. Figure 7 shows examples of automatically generated answers by pre-trained GPT2 and GPT2-FT as well as human authored correct answers. We observe that by fine-tuning on COSMOS, GPT2-FT generates more accurate answers. Although intuitively there may be multiple correct answers to the questions in COSMOS QA, our analysis shows that more than 84% of generated correct answers identified by human are semantically consistent with the gold answers in COSMOS, which demonstrates that COSMOS can also be used as a benchmark for generative commonsense reasoning. Appendix E shows more details and examples for generative evaluation.

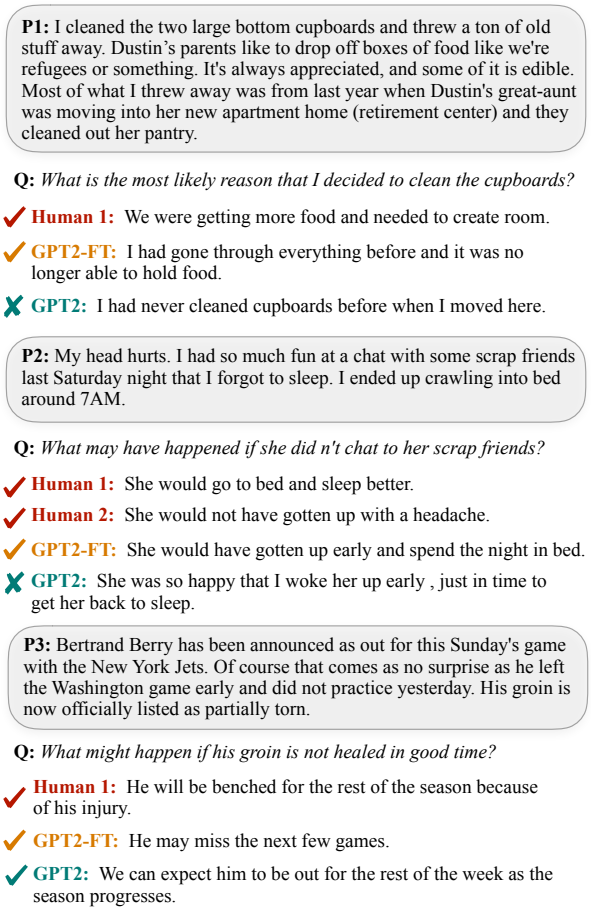


Figure 7: Examples of human authored correct answers, and automatically generated answers by pre-trained GPT2 and GPT2-FT. (✓indicates the answer is correct while ✗shows that the answer is incorrect.)

Metrics	GPT2	GPT2-FT
BLEU (Papineni et al., 2002)	10.7	21.0
METEOR (Banerjee and Lavie, 2005)	7.2	8.6
ROUGE-L (Lin, 2004)	13.9	22.1
CIDEr (Vedantam et al., 2015)	0.05	0.17
BERTScore F1 (Zhang et al., 2019b)	41.9	44.5
Human	11.0%	29.0%

Table 6: Generative performance of pre-trained GPT2 and GPT2-FT on COSMOS QA. All automatic metric scores are averaged from 10 sets of sample output.

6 Related Work

There have been many exciting new datasets developed for reading comprehension, such as SQuAD (Rajpurkar et al.,

Dataset	Size	Type	Answer Type	Paragraph Source	Questions/ Answers	Require MRC	Require Common Sense
MCTest	2K	PQA	MC	MTurk	MTurk	✓	-
RACE	100K	PQA	MC	Human Experts	Human Experts	✓	-
MCScript	13.9K	PQA	MC	MTurk	MTurk	✓	27.4%
NarrativeQA	46.8K	PQA	Open Text	Books/Movie Scripts	MTurk	✓	-
ARC	7.8K	QA	MC	N/A	Web	✗	-
CommonsenseQA	12.2K	QA	MC	N/A	MTurk/Web	✗	100%
ReCoRD	121K	PQA	Span	News	Automatic	✓	75.0%
COSMOS	31.8K	PQA	MC	Webblog	MTurk	✓	93.8%

Table 7: Comparison of the COSMOS QA to other multiple-choice machine reading comprehension datasets: P: contextual paragraph, Q: question, A: answers, MC: Multiple-choice, and - means unknown.

2016), NEWSQA (Trischler et al., 2017), SearchQA (Dunn et al., 2017), NarrativeQA (Kočiskỳ et al., 2018), ProPara (Mishra et al., 2018), CoQA (Reddy et al., 2018), ReCoRD (Zhang et al., 2018), MCTest (Richardson et al., 2013), RACE (Lai et al., 2017), CNN/Daily Mail (Hermann et al., 2015), Children's Book Test (Hill et al., 2015), and MCScript (Ostermann et al., 2018). Most these datasets focus on relatively explicit understanding of the context paragraph, thus a relatively small or unknown fraction of the dataset requires commonsense reasoning, if at all.

A notable exception is ReCoRD (Zhang et al., 2018) that is designed specifically for challenging reading comprehension with commonsense reasoning. COSMOS complements ReCoRD with three unique challenges: (1) our context is from webblogs rather than news, thus requiring commonsense reasoning for *everyday events* rather than *news-worthy events*. (2) All the answers of ReCoRD are contained in the paragraphs and are assumed to be entities. In contrast, in COSMOS, more than 83% of answers are *not* stated in the paragraphs, creating unique modeling challenges. (3) COSMOS can be used for *generative* evaluation in addition to multiple-choice evaluation.

There also have been other datasets focusing specifically on question answering with commonsense, such as CommonsenseQA (Talmor et al., 2018) and Social IQa (Sap et al., 2019), and various other types of commonsense inferences (Levesque et al., 2012; Rahman and Ng, 2012; Gordon, 2016; Rashkin et al., 2018; Roemmele et al., 2011; Mostafazadeh et al., 2017; Zellers et al., 2018). The unique contribution of COSMOS is combining reading comprehension with commonsense reasoning, requiring *contextual* commonsense reasoning over considerably more complex, diverse, and longer context. Table 7 shows comprehensive comparison among the most relevant datasets.

There have been a wide range of attention mechanisms developed for reading comprehension datasets (Hermann et al., 2015; Kadlec et al., 2016; Chen et al., 2016; Dhingra et al., 2017; Seo et al., 2016; Wang et al., 2018b). Our work investigates various state-of-the-art approaches to reading comprehension, and provide empirical insights into the design choices that are the most effective for contextual commonsense reasoning required for COSMOS.

7 Conclusion

We introduced COSMOS QA, a large-scale dataset for machine comprehension with contextual commonsense reasoning. We also presented extensive empirical results comparing various state-of-the-art neural architectures to reading comprehension, and demonstrated a new model variant that leads to the best result. The substantial headroom (25.6%) between the best model performance and human encourages future research on contextual commonsense reasoning.

Acknowledgments

We thank Scott Yih, Dian Yu, Wenpeng Yin, Rowan Zellers for helpful discussions. This research was supported in part by NSF (IIS-1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031), and Allen Institute for AI.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Kevin Burton, Akshay Java, Ian Soboroff, et al. 2009. The icwsm 2009 spinn3r dataset. In *Proceedings of ICWSM 2009*.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the roc story cloze task. In *Proceedings of ACL 2017*, pages 616–622.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of ACL 2016*, pages 2358–2367.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of ACL 2017*, pages 1832–1846.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*.

Andrew S Gordon. 2016. Commonsense interpretation of triangle behavior. In *Proceedings of AAAI 2016*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of NAACL 2018*, pages 107–112.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of NIPS 2015*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of ACL 2018*, pages 328–339.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of ACL 2016*, pages 908–918.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Proceedings of TACL 2018*, pages 317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of EMNLP 2017*, pages 785–794.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of ACL 2017*, pages 510–517.

Bhavana Dalvi Mishra, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. *arXiv preprint arXiv:1805.06975*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51.

Peter Norvig. 1987. *A Unified Theory of Inference for Text Understanding*. Ph.D. thesis, EECS Department, University of California, Berkeley.

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. Mcscript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of LREC 2018*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, page 8.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the winograd schema challenge. In *Proceedings of EMNLP 2012*, pages 777–789.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you dont know: Unanswerable questions for squad. In *Proceedings of ACL 2018*, pages 784–789.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of EMNLP 2016*, pages 2383–2392.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of ACL 2018*, pages 2289–2299.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of EMNLP 2013*, pages 193–203.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2018. Atomic: An atlas of machine commonsense for if-then reasoning. *arXiv preprint arXiv:1811.00146*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of CVPR 2015*, pages 4566–4575.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018a. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 758–762.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018b. A co-matching model for multi-choice reading comprehension. In *Proceedings of ACL 2018*, pages 746–751.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of EMNLP 2018*, pages 93–104.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*.

Shuailiang Zhang, Hai Zhao, Yuwei Wu, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019a. Dual co-matching network for multi-choice reading comprehension. *arXiv preprint arXiv:1901.09381*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu. 2018. Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of AAAI 2018*.