20Q: Overlap-Free World Knowledge Benchmark for Language Models

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, Walter Daelemans
CLiPS Research Center
University of Antwerp, Belgium
maxime.debruyn@uantwerpen.be

Abstract

What do language models know about our world? This question is hard to answer but important to get right. To this end, we introduce 20Q, a novel benchmark using the Twenty Questions game to evaluate world knowledge and common sense of language models. Thanks to our overlap-free benchmark, language models learn the game of Twenty Questions without learning relevant knowledge for the test set. We uncover two intuitive factors influencing the world knowledge of language models: the size of the model and the topic frequency in the pre-training data. Moreover, we show that in-context learning is inefficient for evaluating language models' world knowledge — fine-tuning is necessary to show their true capabilities. Lastly, our results show room for improvement to enhance the world knowledge and common sense of large language models. A potential solution would be to up-sample unfrequent topics in the pre-training of language models.

1 Introduction

Transformers are omnipresent in today's Natural Language Processing. Using a simple training and inference procedure, they reach human-level performance on numerous benchmarks.

The scale of these models is hard to grasp. The most recent one, PaLM (Chowdhery et al., 2022), has 540 billion parameters. It has sixteen times more parameters than all words on Wikipedia, or sixty-eight times more parameters than the total population on Earth (Roser et al., 2013).

Much previous work focused on what these models can do: question-answering, mathematics, translation, or code generation (Wei et al., 2022; Chen et al., 2021; Cobbe et al., 2021; NLLB Team et al., 2022; Lewkowycz et al., 2022). Another exciting area of research is to focus on what these models know: common sense, world knowledge, or biases

	Topic	Question	Answer
🦍	Gorilla	Is it alive?	Yes
⚽	Ball	Can we eat it?	No
⚓	Anchor	Is it heavy?	Yes
✏️	Pen	Can it fly?	No
🚗	Car	Can you drive it?	Yes
🛰️	Satellite	It it furniture?	No

Table 1: Example questions and answers in our 20Q benchmark. We use simple questions to compare the amount of world knowledge between different language models. Despite its apparent simplicity, this benchmark is challenging for even the largest language models — GPT-3 makes a wrong prediction about 20% of the time.

(Kejriwal et al., 2022; Kadavath et al., 2022; Lucy and Bamman, 2021; Abid et al., 2021).

Transformers (Vaswani et al., 2017) models do not store knowledge symbolically — they distribute the knowledge within their weights. As a result, researchers have to use proxy tasks to study it. Previous research used closed-book question-answering datasets to study how much knowledge language models can store (Roberts et al., 2020). They concluded that language models perform similarly with or without external information, thanks to a broad embedded knowledge.

Unfortunately, Lewis et al. (2021) later demonstrated that these datasets suffer from a significant overlap between the training and test set. For example, *who has scored more goals in the premier league* shares the same answer with *most goals scored by a premier league player*. Training on the first and evaluating on the second does not make sense. As a result, T5's (Raffel et al., 2020) performance dramatically dropped when Lewis et al. (2021) removed the overlap – invalidating the conclusion that these models performed equally with or without external knowledge. Our analysis reveals commonsense reasoning benchmarks also display major overlap between the training and test sets.

Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 494 - 508
December 7, 2022 ©2022 Association for Computational Linguistics

Commonsense QA 2.0 Talmor et al. (2022) and Com2sense (Singh et al., 2021) have exact or close-to-exact duplicates between the training and test set.

In this work, we propose a new benchmark, free of any lexical and semantic overlap between the training and test set, to evaluate the world knowledge of large language models using the game of Twenty Questions – a popular yes/no guessing game. See Table 1 for example questions and answers.

We test two hypotheses using this benchmark. First, we test whether large models possess more world knowledge that smaller models. Second, we test our intuition that world knowledge is correlated with the frequency of the topic in language models' pre-training data.

Despite the massive size of GPT-3, it only reaches an F1 score of 82% on our benchmark. It is however much better than its smaller variants, which validates our first hypothesis that larger models possess more world knowledge than smaller models.

Our dataset's unique feature — a generic question and a topic — is ideal for testing our second hypothesis: does world knowledge correlate with topic frequency. Again, the results show our hypothesis is true as the bottom quartile of topics is associated with higher variability, whereas the other quartiles are not.

We conclude this introduction by summarizing our main contributions:

- We release a new benchmark to study the world knowledge of language models. It is free of any overlap between the training and test set.
- We show that large models possess more knowledge than smaller ones. However, the relationship is not linear.
- We show that the knowledgeability of language models on a specific topic depends on the relative frequency of the topic in the pre-training data.

We release our benchmark on the HuggingFace dataset hub (Lhoest et al., 2021) for anyone to use.[1]

[1]https://huggingface.co/datasets/clips/20Q

2 Related Work

Before the rise of deep learning, NLP stored commonsense and world knowledge using semantic networks such as WordNet (Miller, 1995) and later ConceptNet (Speer et al., 2017). These graphs have the advantage of using symbolic representations, facilitating their analysis. Contrary to Transformers-based models, they perform equally well on lower-frequency topics.

Commonsense and world knowledge of Transformers' based models is harder to evaluate, researchers resort to using proxy tasks to evaluate it. Several previous works studied the commonsense abilities of language models in multiple areas: pronoun resolution (Levesque et al., 2012; Sakaguchi et al., 2021), natural language generation (Lin et al., 2020), story understanding (Mostafazadeh et al., 2016), reading comprehension (Zhang et al., 2018; Huang et al., 2019; Ning et al., 2020), physical and social intelligence (Bisk et al., 2020; Sap et al., 2019), temporal reasoning (Zhou et al., 2019), numerical knowledge (Dua et al., 2019; Ravichander et al., 2019), and global commonsense reasoning (Singh et al., 2021; Talmor et al., 2022, 2019).

The remainder of this section focuses on two datasets evaluation the commonsense knowledge of language models using yes/no questions: Commonsense QA 2.0 (Talmor et al., 2022) and Com2Sense (Singh et al., 2021). For both of these datasets, we review the overlap between the training and test set and find troubling examples.

2.1 Commonsense QA 2.0

Talmor et al. (2022) provide a dataset of 14,343 yes/no questions on several commonsense skills: numerical reasoning, causal reasoning, world knowledge, temporal understanding. The authors used a human-in-the-loop approach to create a challenging benchmark for language models. We partially share the same seed data (AllenAI, 2018) as Commonsense QA 2.0, however we follow a stricter pre-processing and split formation procedure.

Overlap Analysis The authors split the training and test sets according to the topic of questions.[2] Our qualitative review of the overlap between the training and test reveals problematic examples. Some examples are almost duplicates: «

[2]For example the question *« an uncle has to have a brother or sister »* has the topic *uncle* even though it also is about the *brother* topic.

an electron holds a positive charge » and, *« an electron holds a positive charge and »*,[3] while others are lexically different but semantically similar: *« most happy meals include a toy »* and, *« happy meals almost always come with a toy »*. We provide more examples in Appendix A.

2.2 Com2sense

Com2sense (Singh et al., 2021) provides a comprehensive commonsense benchmark to test language models' understanding of everyday events and entities by answering yes/no questions. The authors classify their dataset on three axes: knowledge domain (physical, social, or temporal), reasoning scenario (comparative or causal) and numeracy.

Overlap Analysis The authors do not take any special care in the division of the data. However, a key feature of the dataset introduces a high overlap between the two. The authors use a simple technique to double the size of the dataset: edit a few words of each sentence to flip the answer: *to ~~read books~~ see stars at night, one should turn on the lights*. Our qualitative review of the overlap between the training and test reveals highly problematic examples. First, we found exact duplicates between the training and test sets. Second, some examples in the test set are simple negations of examples in the training set. For example *« [...] opening the blinds will help you see »* and, *« [...] opening the blinds will <u>not</u> help you see »*. Third, some examples only change one term between the test and training set, but are semantically similar. We provide more examples in Apppendix A.

2.3 Overlap Analysis Summary

Our qualitative review reveals both of these benchmarks do not properly check for training and test set overlap.

Unfortunately, Lewis et al. (2021) demonstrated that a high overlap between the training and test set can inflate the true performance of language models.

To summarize, we provide the first commonsense reasoning benchmark focused exclusively on world knowledge. Contrary to existing benchmarks, we take extensive measures to ensure there is no overlap between the training and test set. We compare 20Q against alternative benchmarks in Table 2.

[3]the *and* at the end of the sentence is not a typo.

3 Data

Data is a double-edged sword. On the one hand, more data is usually good. However, on the other hand, more data can also complicate the study of the generalization abilities of the model as it gets harder to find uncorrelated validation data.

Regarding world knowledge and common sense, two factors can contaminate the validation data: the training and pre-training data. Large language models can memorize their pre-training data. The bigger the model, the larger the probability of memorization (Chowdhery et al., 2022).

In this work, we take a novel approach and analyze the inner knowledge of large transformers models through the game of Twenty Questions — a popular yes/no guessing game. We take extra care to avoid lexical and semantic overlap between the training and validation sets.

3.1 Twenty Questions Game

Wikipedia describes Twenty Questions as a game that encourages deductive reasoning and creativity. In the traditional game, the answerer chooses a topic and does not reveal it to the questioners, whom themselves must find the hidden entity by asking yes/no questions to the answerer. Humans can play this game (or a variant of it like Guess Who) from a young age.

3.2 Twenty Questions Dataset

We do not generate a dataset ourselves. Instead, we rely on an existing dataset of Twenty Questions games developed by AllenAI, where they had humans play the game of Twenty Questions on Amazon Mechanical Turk. In total, they collected 78,890 questions in the style of Twenty Questions. The dataset is available on Github (AllenAI, 2018).[4]

3.2.1 Generic Questions

As the questioner does not know the topic, he mainly refers to the entity using "it". Therefore, we term these "generic questions." This disentangling of question and topic is helpful in two regards. First, we can use it to ensure no semantic and lexical overlap between the training and validation sets for both topics and questions. Second, we can measure the topic's knowledge by type of word, domain, or relative frequency in the pre-training data.

[4]https://github.com/allenai/twentyquestions

Dataset	Train	Valid.	Test	No Overlap	Focus	Example
CQA2.0	9,264	2,541	2,473	✗	Multiple	*A bus has at least two steering wheels.*
Com2sense	804	402	2,779	✗	Multiple	*As the weather was very cold he put on his jacket to protect himself.*
20Q (ours)	815	-	2,500	✓	World Knowledge	*Can [an acquittal] cheer you up?*

Table 2: Comparison of 20Q with other similar benchmarks. 20Q focuses solely on world-knowledge and is free of any overlap between the training and test set.

3.2.2 Fine-grained Answers

Reducing the world to yes and no can be challenging, even impossible. Instead of answering with yes or no, annotators[5] must answer with fine-grained answers: *never*, *rarely*, *sometimes*, *usually*, or *always*. Three annotators answer each question. With a Kappa score of 57%, the disagreement between annotators is high. However, converting the answers to *yes* or *no* instead of fine-grained answers resolves any disagreement between annotators. Using a binary answer also facilitates the analysis.

3.2.3 Quality Score

Annotators provide a quality score for each question and flag potential problems: questions that are not answerable by yes or no, questions that are not playing the game, or questions that refer to another turn. We only retain questions with the highest quality score (85% of the dataset).

3.3 Pre-processing

As with all data generated by humans, it can be noisy. The original dataset contains many sentences with orthographic errors, or even questions unrelated to the Twenty Questions game. Our goal is to understand the knowledge stored inside the language models, not their capacity to deal with noise. Therefore, we take extensive pre-processing steps to clean the dataset. We give further insight into our pre-processing in Annex B. First, we remove all questions below the maximum score of three (-15%). Next, we remove all questions which do not use "it" (-12%). Finally, we remove all duplicate questions (-3%) and answers where the topic is not in WordNet (-3%). Our pre-processing removes 34% of the initial dataset.

3.4 Training Set

The original authors performed a random split of questions into training, validation, and test set. The authors deal with training/test overlap by flagging questions where the topic is also present in the training set. We take a much stronger stance on train/test overlap and include the semantic overlap between topics and questions.

Our objective is to test the existing knowledge of language models — not to provide new knowledge. Therefore, the priority should be the size of the test set, not the training set. Our training set consists of 815 questions (500 generic questions) on 707 different topics.

3.5 Similarity Metrics

Before removing the overlap between the training and test set, we must first decide which similarity metric to use.

We use three methods to compute the similarity between two topics (words) or questions (sequence of words).

Bag-of-words The simplest method to compare two words or sequences of words is their bag-of-words representations. We first tokenize, remove stop-words, and finally stem the words. This method typically identifies close lexical duplicates such as *is it animal* & *is it an animal*.

WordNet Our second method uses the semantic graph WordNet (Miller, 1995). WordNet excels at identifying synonyms. For example, it will identify that *bike* is a synonym of *bicycle*.

Sentence Transformers Our last method uses Sentence Transformers (Reimers and Gurevych, 2019). It uses pre-trained encoder networks to compute vector representations of sentences (it also works for single words). We can compare the similarity of two sentences (resp. words) by looking at the cosine similarity of their vector representations. We use three different models.

3.6 Test Set

We follow three steps before including an example in the test set:

[5] We want to stress that we are referring to the annotation of the original dataset (AllenAI, 2018).

	Training	Test
Questions (total)	815	2,500
Generic Questions	500	1,250
Topics	707	1,436
Words	5.3	5.2
Yes	46%	42%
No	54%	58%

Table 3: Descriptive statistics. Our goal is not to learn new knowledge but to test existing knowledge. As a result, the training set is small compared to the validation set.

1. We ensure that the bag-of-words representation of the question and the topic is not present in the training set.
2. We check if the topic of the question is not a synonym of any topic in the training set.
3. Our last step removes any example with a cosine similarity larger than 0.8 with any topic or question in the training set.

After all these steps, we arrive at a test set of 4,201 examples. Given the high cost of evaluating very large language models, we only keep the first 2,500 examples. Given the limited size of the validation set, we did not implement a test set. Additional statistics about the dataset are available in Table 3. Our validation consists of only 4% of the clean dataset. However, as there is no overlap between the training and validation set, we can make safe conclusions on the generalization abilities of language models.

4 Overlap Exploration

Lewis et al. (2021) demonstrated the devastating effect of an uncontrolled overlap between the training and validation set. Therefore, this section uses different techniques to inspect the most similar items between the training and validation set.

4.1 Topic Overlap in 20Q

We start by analyzing the overlap in topics. For example, we want to avoid having questions about *cars* in the training set and about *automobiles* in the validation set.

N-grams Character n-grams are a good way to retrieve words sharing almost the same lexical form.[6]

[6]We use a character tri-grams

We show the five most similar pairs of topics between the training and validation set in Table 7 in Annex C. The most similar topics according to this method are *account* and *accountant*. This technique does not reveal problematic overlap between the two sets.

WordNet We use WordNet to compute the distance between two topics by following the hypernym or hyponym chain. Table 8 in Annex C shows this technique's most similar pair of topics. None of the retrieved pairs show a significant semantical or lexical overlap.

Sentence Transformers We finish our qualitative review of the topic overlap using Sentence Transformers. Table 9 in Annex C shows the five most similar pairs of topics. The most similar pairs are *costume* with *halloween*, *chlorophyll* and *chrysanthemum*, *bracelet* and *pendant*. All of these words are related, but none are synonyms of one another.

4.2 Question Overlap in 20Q

An overlap in terms of topics is only part of the story. We also want to avoid evaluating models on the same kind of answers used to train them. Therefore, we perform the same procedure to avoid lexical and semantic overlap between the questions in the training and validation set. The task is trickier than for topics. For example, *Does it make you cry* and *Does it make you laugh* only differ in a single token, but their meaning is opposite.

BM25 We use BM25 to retrieve similar questions between the two sets. The two most similar questions are *Can the human population fit on it?* and *Would it fit in the palm of a human hand?*. These questions share two important tokens: *fit* and *human*, but they do not have the same meaning. See table 4 for more examples. This clearly shows how semantically inequivalent even the most similar sentences in the train and validation set are.

Sentence Transformers Next, we perform the same analysis with Sentence Transformers. The most similar questions between the two sets are *does it have a steering wheel?* and *does it have gears or screws?*, indicating a sufficient amount of dissimilarity between the questions in the training and test set.

Train	Topic	Validation	Topic
Does it have a one time function?	knocker	Does it need to be one student at a time?	lettering
Would a parent want their child to do it?	soloist	Is it a category response, like parent or child?	cornea
Can the human population fit on it?	earth	Would it fit in the palm of a human hand?	keyboard
Does it rock?	brim	Is it some sort of precious, rare stone or rock?	emerald
Is it a turn?	heron	Is it something you turn on?	dice

Table 4: Qualitative review of the most similar pair of questions computed using BM25. Questions usually share a similar word (e.g., *child* or *rock*), however, it is used in a different context each time. Moreover, the topics are completely unrelated, reducing the risk of overlap even more.

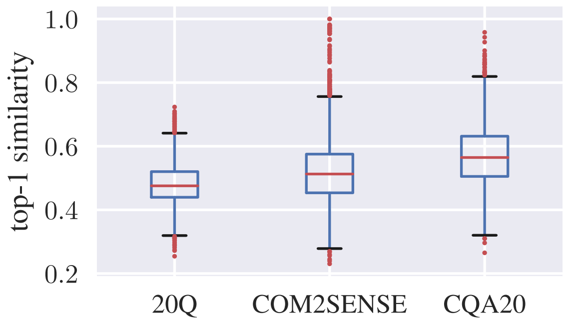

Figure 1: Distribution of top-1 similarity between examples in the training and test set. 20Q has the lowest similarity between the two (by design).

4.3 Comparison with Existing Benchmarks

We finish this section by comparing the train/test overlap of 20Q with two existing benchmarks presented in Section 2: Commonsense QA 2.0 and Com2sense. For each question in the test set, we look for the most similar one in the training set using Sentence Transformers. We summarize the results in Figure 1. The results are striking, 20Q has significantly less overlap with the training set than Com2sense and Commonsense QA 2.0. Our qualitative analysis of these results reveal dangerously close duplicates between the training and test of these two benchmarks. Even less expected, we uncover exact duplicates between the training and test of Com2sense. We provide a more detailed analysis in Annex A.

To summarize, our benchmark is free of any semantic and lexical overlap between the training and validation set regarding topics and questions. Moreover, despite the strict separation constraints, both sets stay semantically diverse.

5 Language Model

After reviewing that data, we review the language models. Although previous work used text-to-text models such as T5 (Raffel et al., 2020), T0 (Sanh et al., 2022), and BART (Lewis et al., 2020), in this work, we stick to GPT-3 (Brown et al., 2020), a general-purpose decoder-only Transformers language model. By sticking to a single model, we can ensure that the only differentiating factor between the models is the network size, not the pre-training data or model architecture.

5.1 GPT-3

GPT-3 (Brown et al., 2020) is an auto-regressive language model developed by OpenAI. The model weights are not publicly available, although the model's predictions are available through a paid API.

Size GPT-3 comes in four sizes: 2.7B, 6.7B, 13B and 175B. We use this feature to understand how the size of a model influences the amount of world knowledge it can store.

Pre-training Data The authors of GPT-3 did not release the pre-training data used to train the model. So instead, we use C4, the dataset used to train T5 (Raffel et al., 2019), as a proxy to estimate the frequency of each topic in our benchmark.

Prompting GPT-3 was never trained to answer yes/no questions. Instead, its objective is to predict the next token in a piece of text. The standard way to query a large language model is to use in-context learning, where one provides a few examples of the task in the prompt and asks the language model to complete the last example.

6 Experiments

Our experiments aim at understanding which models possess the best world knowledge. We believe large language models are ineffective at querying their internal knowledge using in-context learning. For this reason, we also fine-tune each model on the training set for a single epoch. The goal is not to teach new knowledge but to guide the model into learning the task. As we meticulously assembled

Model	Size	F1 Z-S	F1 F-S	F1 F-T	NLL Z-S	NLL F-S	NLL F-T
GPT-3	2.7B	58.77	58.02	58.04	112.9	82.64	66.46
GPT-3	6.7B	58.45	54.53	66.35	140.5	80.56	55.41
GPT-3	13B	59.65	48.88	74.48	79.87	65.52	55.63
GPT-3	175B	61.10	67.14	**82.50**	69.86	62.23	**41.16**

Table 5: Results per model size and inference method: zero-shot (Z-S), few-shot (F-S), and fine-tune (F-T). According to F1 and NLL, the best method is the largest GPT-3 fine-tuned on our training set.

our training and validation splits, we are sure any performance gain will not come from the knowledge acquired during fine-tuning.

6.1 Zero-shot

The zero-shot approach is the simplest way to evaluate the knowledge of the language model. The model must predict the next token without any prior examples. We record the probability of the yes token and no token.

Prompt

```
You are playing a game of 20 questions.
Answer the following question
about with yes or no.

Topic: {{ question_topic_1 }}
Question: {{ question_example_1 }}
Answer:
```

6.2 Few-shot

This approach improves upon the previous one by providing multiple examples to steer the model in the right direction. The model learns the task *on the fly* using examples from the training set. We record the probability of the yes token and no token.

Prompt

```
Topic: {{ topic_example_1 }}
Question: {{ question_example_1 }}
Answer: {{ answer_example_1 }}
...
Topic: {{ topic_example_n }}
Question: {{ question_example_n }}
Answer:
```

Settings We provide four examples in a random order (two positives and two negatives) from the training set.

6.3 Fine-tuning

Understanding the task of answering yes/no questions using on the fly examples is hard. Therefore, we also tested another approach where we fine-tuned models on our training set.

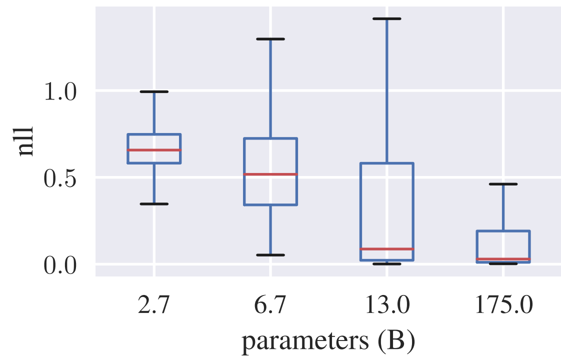

Figure 2: Box-plot of negative-likelihood (NLL) per model size.

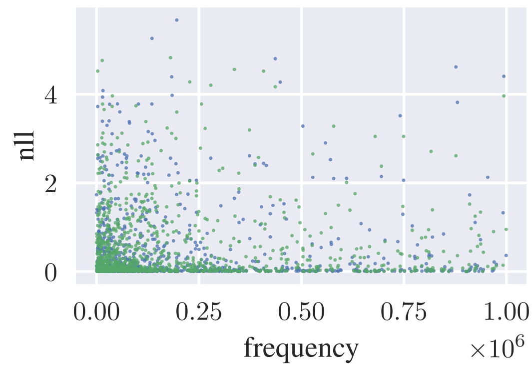

Figure 3: Scatter plot of NLL by topic frequency for the 13B (blue) and 175B (green) models.

Prompt

```
Topic: {{ topic_example }}
Question: {{ question_example }}
Answer:
```

Settings Each model is trained on a single epoch of the training set.

7 Results

We run all experiments and report binary-F1 and Negative Log-Likelihood (NLL) to the ground-truth answers in Table 5. We start by reviewing the effect of fine-tuning and then analyze our two hypotheses.

7.1 Fine-tuning

The benefit of fine-tuning is clear: fine-tuned models are systematically better than few-shot and zero-shot across model size and evaluation metrics. Moreover, thanks to our detailed review of the overlap, we can safely assume the out-performance does not come from learning any new knowledge but is due to better use of the world knowledge already present in the language models.

7.2 Size Effect

In theory, the larger the model, the more space it has to store world knowledge. Therefore, we expect to see better performance for large models. Figure 2 shows a box-plot of the negative log-likelihood of the fine-tuned results by the model size.

The results are somewhat unexpected. Although the median negative log-likelihood is steadily declining with the model size, the variability also increases with the model size, except for the largest one, which breaks the trend with a low median loss and low variability. In other words, the model's ability to know what it does not know diminishes with model size.

7.3 Frequency Effect

Previous research showed that the frequency of tokens in the pre-training data influences the ability of large language models to do numeric reasoning (Razeghi et al., 2022). We hypothesize that the same is true when it comes to world knowledge. Language models should have a harder time answering questions on topics they have rarely encountered during pre-training. Therefore, we collected the frequency count of each topic in a large pre-training corpus: C4 (Raffel et al., 2020). Our experiments revealed the high correlation of topic frequency with the perplexity of GPT-2 (XL) to generate the word. We use this metric as it scales to different word forms and is easier to collect. [7]

Figure 3 clearly shows the frequency effect. Topics associated with a lower frequency quartile have more variability in negative log-likelihood than higher quartiles. This effect is especially strong on the 13B model.

7.4 Question Bias

In this section, we try to uncover whether language models use statistical cues in the question rather than their internal knowledge to answer questions. To this end, we run the fine-tuned model (explained in Section 6.3) without the topic in the prompt. If language models use statistical patterns in questions, it should not matter whether the subject is present or not. The F1 score of GPT-3 (175B) drops from 82.50% to 59.40%, just over the performance of the smallest GPT-3 model. We conclude that language models use their internal knowledge rather than statistical cues in the questions.

[7] We use the cross-entropy loss (using a sum reduction) from a GPT-2 XL model as a measure of frequency

8 Conclusion

Previous research (Lewis et al., 2021) showed that language models do not have enough world knowledge to rival open-domain question-answering systems. We update this claim using larger models and a novel benchmark, 20Q. We find two factors influencing the world knowledge of language models: the model's size and the topic's frequency in the pre-training data. Thanks to careful attention to the overlap between the training and validation set, we can safely conclude that fine-tuning provides a better picture of the world knowledge possessed by language models. Our benchmark shows that even the largest language models (175 billion parameters) have room for improvement regarding world knowledge. We propose several areas of improvement for coping with a rapidly changing world as future work.

Acknowledgement

We thank the reviewers for their helpful feedback. This research received funding from the Flemish Government under the *Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen* programme.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 298–306, New York, NY, USA. Association for Computing Machinery.

AllenAI. 2018. A web application for playing 20 questions to crowdsource common sense.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Mayank Kejriwal, Henrique Santos, Alice M Mulvehill, and Deborah L McGuinness. 2022. Designing a strong test for measuring true common-sense reasoning. *Nature Machine Intelligence*, 4(4):318–322.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

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

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55, Virtual. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

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

Abhilasha Ravichander, Aakanksha Naik, Carolyn Rose, and Eduard Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 349–361, Hong Kong, China. Association for Computational Linguistics.

Yasaman Razeghi, Robert L. Logan, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Max Roser, Hannah Ritchie, and Esteban Ortiz-Ospina. 2013. World population growth. *Our World in Data*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine

Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of AI through gamification. *ArXiv*, abs/2201.05320.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

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*, abs/1810.12885.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

A Detailed Overlap Analysis

In this section, we review the most similar pairs of questions between the training and test for Commonsense QA 2.0, Com2sense, and 20Q (our benchmark). We use Sentence Transformers (Reimers and Gurevych, 2019) to compute the similarity between all pairs of questions in the training and test set.

A.1 Commonsense QA 2.0

The authors of Commonsense QA 2.0 used a topical split to divide the training and test set. We list the top 15 most overlapped questions between the training and test set in Table 11. A quick analysis of the table reveals a number of problematic pairs such as *« an electron holds a positive charge and »* is an almost duplicate to *« an electron hold a positive charge »*.

A.2 Com2sense

Our overlap analysis of com2sense reveals three *exact duplicates* between the training and test set of Com2sense. A number of examples are close duplicates and only change with one word or punctuation. For example *« if it is dark outside, opening the blinds will not help you see »* and *« if it is dark outside opening the blinds will help you see »*. We list the top fifteen overlap pairs in Table 12.

A.3 20Q

Our overlap analysis of 20Q does not reveal any overlap thanks to our strict pre-processing pipeline. We list the top fifteen overlap pairs in Table 10.

A.3.1 UMAP

Figure 4 and 5 provide a 2 dimension projection of the semantic of questions and subject in 20Q.

B Pre-processing

The original Twenty Questions dataset is generated by humans, and is thus extremely noisy. In this section, we expand upon Section 3.3 and go into the details of our pre-processing steps. We detail our pre-processing steps and the percentage of questions removed in Table 6.

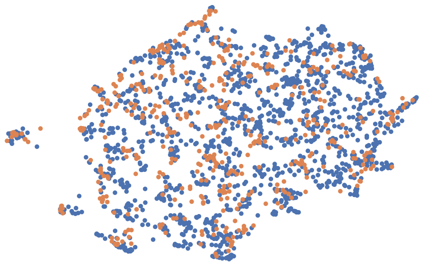

Figure 4: UMAP projection of the Sentence Transformers representation of the questions. Blue dots belong to the training set, red dots belong to the validation set.

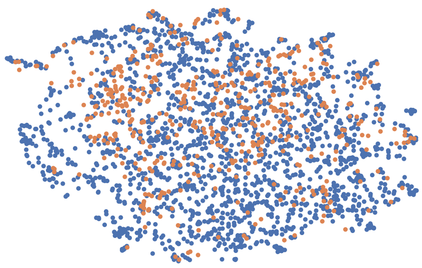

Figure 5: UMAP projection of the Sentence Transformers representation of the topics. Blue dots belong to the training set. Red dots belong to the validation set.

Step	Size (abs)	Size (%)
Initial dataset	78,890	100
Low scores	-12,396	-15.7
Do not use "it"	-9,665	-12.3
Duplicates	-2,708	-3.4
WordNet	-2,312	-2.9
Clean dataset	51,809	65.7

Table 6: Pre-processing of the original dataset. We are aggressive in our pre-processing as we prefer a small dataset of high quality to the reverse. First, we remove all questions with a score of 2 (the maximum is 3). We then remove all sentences that do not use "it." Next, we use a stemmed bag-of-words representation to remove close duplicates. Finally, we remove all questions where the answer is not in WordNet.

B.1 Quality Score

We start our pre-processing by removing all sentences with a score below three. These are questions which are not answerable with *yes* or *no*, or questions which are not playing the game of Twenty Questions. For example, questions such as *« so not an object, but tangible. is it edible »* which references the previous turn, or simple one word questions such as *« mountain? »*

B.2 Use of *it*

Our goal is to understand the world knowledge of language models. For some models such as T0 or T5, it may be easier to answer the question if the topic is part of the question, instead of having two separated parts. For example it is easier to answer: *« does a rock float »* than *« subject: rock, question: does it float »*. To make sure all questions are equally easy or difficult in terms of lexical information, we only keep questions of the latter format.

B.3 Duplicate Questions

Some questions may be close, but not exact, duplicates. We want to avoid such questions in the training or test set as these add very little information while artificially inflating the size of the dataset. We use a stemmed bag-of-words approach to detect these questions. For example, questions such as *« is it animal »* and *« is it an animal »*.

B.4 WordNet Filtering

We want to avoid having questions where the subject is not orthographically correct. We remove all questions where the subject is not present within WordNet. In effect, this will remove words such as *trex*, *chldren*, *voiceing*, or acronym words such as *potus* or *49ers*.

C Topic Overlap Exploration

In this section, we show the list the overlapping topics according to three different metrics.

C.1 N-grams

We show the five most similar pairs of topics between the training and validation set in Table 7.

C.2 WordNet

We use WordNet to compute the distance between two topics by following the hypernym or hyponym chain. Table 8 shows this technique's most similar pair of topics.

Train	Validation	Sim.
Account	Accountant	0.84
Thinking	Thing	0.79
Constitution	Institution	0.78
Extraction	Traction	0.78
Attraction	Traction	0.78

Table 7: Most similar pair of topics between the training and validation set using a character tri-gram method.

Train	Validation	Sim.
Vegetation	Galaxy	0.33
Purifier	Pendulum	0.33
Lambskin	Squirrel	0.33
Foil	Steel	0.33
Repellent	Menthol	0.33

Table 8: Most similar pair of topics between the training and validation set using the WordNet method.

C.3 Sentence Transformers

We finish our qualitative review of the topic overlap using Sentence Transformers. Table 9 shows the five most similar pairs of topics.

Train	Validation	Sim.
Costume	Halloween	0.60
Chlorophyll	Chrysanthemum	0.60
Housekeeper	Groomsman	0.60
Bracelet	Pendant	0.60
Forearm	Ankle	0.60

Table 9: Most similar pair of topics between the training and validation set using the Sentence Transformers method.

Test Set	**Training Set**
would it [a granite] be of rock material?	can it [a rock] be molded?
is it [a window] see through?	does it [a curtain] cover a window?
is it [a sweat] produced by the human body?	does it [an exercise] involve sweating?
does it [a hyacinth] have red flowers?	does it [a chrysanthemum] have a long stem?
is it [a ring] jewlery?	does it [a treasure] go on engagement rings?
is it [a bridge] larger than a car?	is it [a bumper] a bridge?
is it [a refuge] a type of campsite?	is it [a campground] the mountains?
is it [an ant] bigger than a honeybee?	does it [a honeybee] collect nectar?
is it [a marsupial] a kind of bear?	is it [a bear] long?
does it [a hyacinth] have white flowers?	does it [a chrysanthemum] have a long stem?
is it [a pendant] jeweled?	does it [a treasure] go on engagement rings?
does it [a hyacinth] have yellow flowers?	does it [a chrysanthemum] have a long stem?
is it [a ship] larger than a whale?	does it [a whale] have fins?
is it [a hurdle] made of stone or rock?	can it [a rock] be molded?
is it [a fly] a bug?	does it [an insect] have antennae?

Table 10: Top fifteen most similar pairs of questions between the training and test set of 20Q.

Test Set	**Training Set**
an electron holds a positive charge and	an electron holds a positive charge.
happy meals almost always come with a toy.	most happy meals include a toy.
april is larger than february	april is smaller than march
sunlight on the skin causes eye cancer	sunlight causes almost all skin cancer
thunder sounds before lightning strikes	noise of thunder is heard before the lightning.
the beginning of a story is part of the end	a story has a beginning and an end.
is there a feminine french word for a city hall?	in french is it true that there are feminine and masculine words for a city hall?
europe is considered to be the most wealthy and richest continent.	europe has the richest countries in the world
a grapefruit is a fruit larger than a watermelon?	is a watermelon smaller than an apple?
tree is always part of forest	trees are never part of forests
someone of the male gender cannot give birth.	an adult male cannot give birth
if you add two plus two you will always get four.	two plus two unfortunately cannot ever add up to anything but four.
you can return items to a store only if you have a receipt.	an item can be returned from a store only if it is sold by that store.
private is another way to say public	private almost never means public.
a letter can be written with invisible ink.	writing cannot be read if you use invisible ink.

Table 11: Top fifteen most similar pairs of questions between the training and test set of Commonsense QA 2.0.

Test Set	Training Set
john leaves work at 6 pm so that he is an unlikely suspect for theft that happened in the office at 8 pm.	john leaves work at 6 pm so that he is an unlikely suspect for theft that happened in the office at 8 pm.
while in a windy rainstorm, you should always point your umbrella away from the wind.	while in a windy rainstorm, you should always point your umbrella away from the wind.
while in a windy rainstorm, you should always point your umbrella into the wind.	while in a windy rainstorm, you should always point your umbrella into the wind.
since i want to improve my golf skill quickly, i spend 2 hours on the course every day.	since i want to improve my golf game, i spend 2 hours on the course every day.
if it is dark outside, opening the blinds will help you see.	if it is dark outside opening the blinds will not help you see.
because it was halloween eve and we had no candy, i decided to open the door and turn the porch light on.	because it was 6pm on halloween and we no candy, i decided to open the door and turn the porch light on.
having to teach a night class in thirty minutes, he should cook a three-course dinner instead of heating a frozen meal.	having to teach a night class in thirty minutes, he should make a three-course dinner instead of a frozen meal.
danny smokes a lot and drinks thirty beers per week while sarah doesn't smoke and doesn't drink, sarah will probably live longer.	danny smoke a lot and drink thirty beer per week while sarah dont smoke and dont drink, sarah will probably live longer.
if it is dark outside, opening the blinds will not help you see.	if it is dark outside opening the blinds will not help you see.
because it was halloween eve and we had plenty of candy, i decided to open the door and turn the porch light on.	because it was 6pm on halloween and we had plenty of candy, i decided to open the door and turn the porch light on.
having to teach a night class in thirty minutes, he should heat a frozen meal instead of cooking a three-course dinner.	having to teach a night class in thirty minutes, he should make a frozen meal instead of a three-course dinner.
danny smokes a lot and drinks thirty beers per week while sarah doesn't smoke and doesn't drink, danny will probably live longer.	danny smoke a lot and drink thirty beer per week while sarah dont smoke and dont drink, danny will probably live longer.
a spoon is more suitable for eating soup than a fork.	a spoon might be more suitable for eating soup than a fork.
it is easier to run one mile in 5 minutes than a half mile in 10 minutes.	it is easier to run two miles in five minutes than it is to run one mile in ten minutes.
a fork is more suitable for eating soup than a spoon.	a spoon might be more suitable for eating soup than a fork.

Table 12: Top fifteen most similar pairs of questions between the training and test set of Com2sense.