Can NLP Models 'Identify', 'Distinguish', and 'Justify' Questions that Don't have a Definitive Answer?

Ayushi Agarwal* **Nisarg Patel*** **Neeraj Varshney*** **Mihir Parmar**
Pavan Mallina **Aryan Shah** **Srihari Raju Sangaraju**
Tirth Patel **Nihar Thakkar** **Chitta Baral**
Arizona State University

Abstract

Though state-of-the-art (SOTA) NLP systems have achieved remarkable performance on a variety of language understanding tasks, they primarily focus on questions that have a correct and a definitive answer. However, in real-world applications, users often ask questions that don't have a definitive answer such as questions about future events, questions lacking necessary details to find the answer, and questions that are ambiguous. Incorrectly answering such questions certainly hampers a system's reliability and trustworthiness. Can SOTA models accurately identify such questions and provide a reasonable response?

To investigate the above question, we introduce `QnotA`, a dataset consisting of five different categories of questions that don't have definitive answers. Furthermore, for each `QnotA` instance, we also provide a corresponding `QA` instance i.e. an alternate question that "*can be*" answered. With this data, we formulate three evaluation tasks that test a system's ability to '*identify*', '*distinguish*', and '*justify*' `QnotA` questions. Through comprehensive experiments, we show that even SOTA models including GPT-3 and Flan T5 do not fare well on these tasks and lack considerably behind the human performance baseline. We conduct a thorough analysis which further leads to several interesting findings such as, despite not being able to accurately identify a QnotA question, GPT-3 on being prompted to output a justification of why the given QnotA question doesn't have a definitive answer is able to provide a reasonable justification. Finally, we believe our work and findings will encourage and facilitate development of more robust NLP systems that can also reasonably respond to questions that don't have a definitive answer.

1 Introduction

Recent advancements in Natural Language Processing (NLP) have led to the development of Question-Answering (QA) systems that possess remarkable capabilities of providing fluent and comprehensive answers to a wide range of questions (Khashabi et al., 2020; Brown et al., 2020; Zhang et al., 2022; Lourie et al., 2021; Chowdhery et al., 2022; Rae et al., 2021). However, most of these systems focus on questions that have a correct and an objective answer. But, users in real-world applications often ask questions that are either about some future event, lack the necessary details to reach to a conclusion, or are factually incorrect; essentially, questions that don't have definitive answers. Incorrectly answering such questions can have serious consequences and can thus hamper the system's reliability and trustworthiness. Can state-of-the-art NLP models accurately identify such questions and provide a reasonable response?

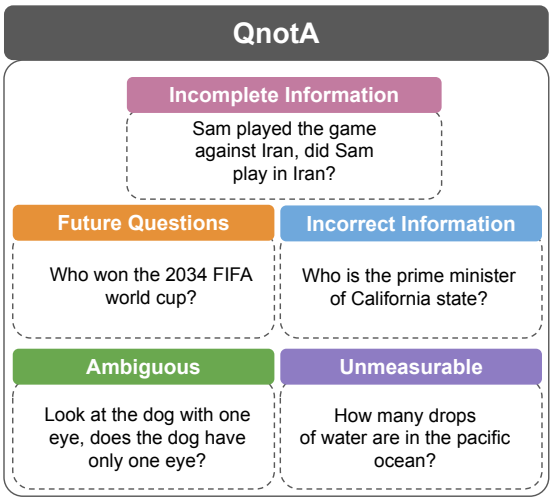

Figure 1: Illustrative examples of different categories of questions from `QnotA`.

In this paper, we investigate the above question. Specifically, we first introduce `QnotA`, a dataset consisting of five different categories of questions that don't have a definitive answer. Figure 1 illustrates examples of all the categories. Furthermore, for each `QnotA` instance, we also provide a cor-

*Equal Contribution

Category	QnotA	QA
Ambiguous	The lecturer said on Friday she would take a pop quiz. When is the pop quiz?	The lecturer said that she would take a pop quiz on Friday. When is the pop quiz?
	Look at the dog with one eye. Does the dog have only one eye?	Look at the one-eyed dog. Does the dog have only one eye?
	Sam went for a walk with his friend in the red shirt. Who was wearing the red shirt?	Sam went for a walk with his friend who was in the red shirt. Who was wearing the red shirt?
Incomplete Information	The band released a new album on life in Paris. Did the band release the new album in Paris?	The band released a new album in Paris on the topic 'life in Paris'. Did the band release the new album in Paris?
	Lisa is planning a vacation abroad. What country is she planning to visit?	Lisa is planning a vacation to Japan. What country is she planning to visit?
Incorrect Information	What animal can be found at the top of the men's Wimbledon trophy?	What fruit can be found at the top of the men's Wimbledon trophy?
	When did Lebron James start playing cricket?	When did Lebron James start playing basketball?
	When did Edmund Hillary climb the Gangkhar Puensum?	When did Edmund Hillary climb the Mount Everest?
Future Questions	Who won the 2034 FIFA world cup?	Who won the 2018 FIFA world cup?
	Who won the presidential election in 2040?	Who won the presidential election in 2016?
Unmeasurable	What is the total weight of all the ants?	What is the average weight of an ant?
	Where can we find heaven?	Where can we find books on heaven?

Table 1: Examples of QnotA and corresponding QA instances for all categories in our dataset.

responding QA instance i.e. an alternate question that "*can be*" answered. Then, using this data, we formulate the following three evaluation tasks: (a) given a question, **identify** whether it has a definitive answer, (b) given two questions (QnotA and its corresponding QA instance, not in order), **distinguish** which one has a definitive answer, and (c) given a QnotA instance, **justify** why it does not have a definitive answer.

We conduct comprehensive experiments with several state-of-the-art models such as GPT-3 (Brown et al., 2020), Flan T5 (Chung et al., 2022), and also models fine-tuned on standard NLP datasets. We first show that even state-of-the-art models achieve considerably lower performance than the human performance baseline. We also explore the impact of providing in-context examples and re-framing instructions in the prompt. We show that it helps reduce the performance gap between model and human performance. However, there is still a considerable performance difference, implying the scope of future research in this direction. For the third task, we give a QnotA instance and prompt the system to justify why it does not have a definitive answer. We find that despite not being able to accurately identify a QnotA question, GPT-3 on being prompted to output a justification of why the given QnotA question doesn't have a definitive answer is able to provide a reasonable justification.

Given the practical utility of data created in this work and the fact that human-generated data is often expensive and time-consuming to collect, we also explore expanding this data using automated means. To this end, we use the generation capabilities of GPT-3 model to expand the dataset. We also check the synthetically created questions and show their validity via a validation step.
In summary, our contributions are as follows:

1. To investigate the ability of NLP models to appropriately respond to questions that don't have definitive answers, we introduce QnotA, a dataset consisting of five diverse and different categories of questions that don't have definitive answers.
2. We formulate three evaluation tasks that eval-

uate a system's ability to 'identify', 'distinguish', and 'justify' QnotA instances.
3. We show that even SOTA models like GPT-3 and Flan T5 achieve considerably lower performance than humans. We also conduct a thorough analysis which further leads to several interesting findings.
4. We also explore creating more such data instances which can be of practical significance for further research.

Overall, we believe our work will encourage and facilitate further research in this important area and help develop more reliable NLP systems that can enable their safer, trustworthy, and widespread adoption in real-world applications.

2 Related Work

Recently, numerous datasets have been created that test different language understanding skills such as pronoun resolution (Sakaguchi et al., 2021; Levesque et al., 2012), numerical reasoning (Ravichander et al., 2019; Lin et al., 2020; Zhang et al., 2020; Mishra et al., 2022b), commonsense reasoning (Singh et al., 2021), qualitative reasoning (Tafjord et al., 2019), physical commonsense reasoning (Bisk et al., 2020), temporal (Zhou et al., 2019), and feasibility understanding (Gupta et al., 2023). Furthermore, other datasets that include questions with false presuppositions (Kim et al., 2021, 2022), ethical risks (Weidinger et al., 2021), and ambiguous external knowledge (Min et al., 2020) also have been proposed. Despite having practical significance, language understanding skill corresponding to identifying, distinguishing, and justifying questions that don't have definitive answers has remained underexplored. Existing datasets lack an ample number of such instances to evaluate models on this crucial skill. In this work, we introduce a collection of five different categories of such questions, formulate three evaluation tasks, and conduct a thorough investigation with several state-of-the-art models.

'Selective prediction' or 'rejection' task in which a model can decide to abstain from prediction when it is likely to be incorrect is related to our work. Selective prediction has previously been studied for abstaining on questions that are either difficult (when the model is uncertain) or out-of-distribution or novel (Kamath et al., 2020; Varshney et al., 2022a,b; Xin et al., 2021; Xu et al., 2022; Varshney and Baral, 2023; Kadavath et al., 2022; Varshney et al., 2023). In this work, we focus on questions that don't have definitive answers.

3 QnotA

QnotA consists of questions that do not have a definitive answer. Furthermore, for each QnotA instance, we also provide a corresponding QA instance i.e. an alternate question that "can be" answered. All our data instances are in the English language. Nine computer science graduate students contributed towards the creation of this dataset and are also part of the author list of this paper. Our dataset consists of five different categories as detailed below:

3.1 Dataset Categories

Incomplete Information: Consider the question, "David played his last match against Australia, did David play his last match **in Australia**?" The context in this question lacks information about the location of the game and hence, the question 'did he play in Australia' can not be definitively answered. In this category, we include such questions that fail to provide sufficient information to reach to a conclusion. An ideal response to such questions should highlight the lack of necessary information in the context.

Future Questions: This category includes questions about things and events that are yet to happen and their outcome can not be exactly predicted. For example, "Who won the presidential election in 2040?" It is clear that such questions do not have a definitive correct answer. An ideal response to the above question should highlight that the 2040 election has not happened yet and thus can't be definitively answered.

Incorrect Information: Consider the question, "When did Italy invade China", the question mentions incorrect information as Italy never invaded China. An ideal system should provide an appropriate response highlighting the flaw in the question. For the aforementioned question, a reasonable response could be 'Italy never invaded China'.

Ambiguous: This category includes questions that can be interpreted in multiple ways. Different ways of interpretation can lead to different conclusions, therefore, such questions do not have a definitive answer. For example, "Look at the dog with one eye. Does the dog have only one eye?"

A reasonable response to such a question can mention different interpretations along with their predictions or can even ask for clarifications instead of assuming one interpretation.

Unmeasurable: Consider the question, "How many drops of water are in the sea?" Though in theory, the answer to this question can be quantified, it is still unmeasurable and is not definitive.

Table 1 shows examples of QnotA and corresponding QA instances for all categories in our dataset. We note that in our dataset, we cover a diverse set of questions across the above five categories that test models' ability to identify, distinguish, and justify the questions that do not have definitive answers. **It is in no way an exhaustive list and can thus be further extended with more categories of questions in the future.**

3.2 Dataset Statistics

For each category in `QnotA`, we create 40 human-authored questions with a justification of why each question does not have a definitive answer. In addition, for each QnotA question, we provide an alternate QA question that *can be* answered. Hence, QnotA consists of 200 human-authored question pairs (400 questions) in total. We conduct our primary investigations using these high-quality human-authored examples. However, besides these 400 manually created QnotA instances, we also explore scaling up this dataset by synthetically creating questions using the generation capabilities of GPT-3 (Brown et al., 2020). We study the validity/correctness of human-authored instances in section 3.3 and synthetically created questions in section 5. We also discuss the utility of synthetically generated questions in Section 5. Finally, we will release our dataset at <anonymous link>.

3.3 Dataset Validation

The human-authored data instances were cross-verified by three other students and the instances where the inter-annotator agreement was low were rejected. A total of six instances were rejected/ modified in this validation step. Furthermore, we also cross-verified the categories assigned to the data instances.

4 Experiments and Results

4.1 Experimental Setup

Let a QnotA question be denoted by q' and its corresponding alternate QA question by q then (q', q) is a question pair in which the question q' doesn't have a definitive answer while q can be answered.

Task 1: In the first task, a question is given as input and the system needs to identify whether it has a definitive answer or not. Here, the system is presented with one question at a time from both QnotA and QA sets (combined and shuffled), and the performance is measured based on how often it correctly identifies whether the given question has a definitive answer or not. Notice that having paired data instances q' and q further allows us to evaluate the system's 'consistency' in its predictions. We define consistency as the fraction of correctly predicted question pairs i.e. we measure the fraction of question pairs in which q' is predicted as 'does not have a definitive answer' and q is predicted as 'has a definitive answer'.

Task 2: Here, we provide two questions (QnotA and its corresponding QA instance, not in order) as input and the system is required to identify which one has a definitive answer. Note that it is different from measuring consistency in Task 1 because here the system needs to select one of two questions as the question having a definite answer while in Task 1, the two instances are independently inferred and the system can predict any label for each instance.

Task 3: In the third task, a QnotA instance is given as input and the system needs to justify the reason why the given question does not have a definitive answer. Since this is not a classification task, we conduct human evaluations for evaluating the correctness of model predictions.

4.2 Human Performance Baseline

We collect a total of 3 responses from different individuals for each instance and use the majority voting aggregation method to measure the human performance baseline for all the tasks. Measuring performance for tasks 1 and 2 is straightforward (calculate accuracy with the ground truth labels). However, for task 3, the ground truth is not a label but a sentence justifying the reason why the given question does not have a definitive answer. Hence, for this task, the authors measure the performance by marking each prediction as correct or incorrect. We note that the same evaluation methodology is used for model evaluations also.

Category	Human	GPT-3	FlanT5	Bart-MNLI
Incomplete Information	97.29	$88.00_{2.61}$	$84.75_{3.52}$	$49.00_{1.79}$
Future Questions	72.50	$53.00_{4.06}$	$50.00_{0.44}$	$50.25_{0.45}$
Incorrect Information	70.00	$71.00_{10.26}$	$50.00_{0.00}$	$50.00_{0.00}$
Ambiguous	83.54	$55.60_{4.72}$	$53.50_{5.15}$	$49.50_{0.62}$
Unmeasurable	72.61	$70.30_{8.76}$	$51.00_{1.86}$	$49.25_{1.79}$
Avg	79.18	$67.58_{6.08}$	$57.85_{2.19}$	$49.60_{0.93}$

Table 2: Accuracy achieved by different models on Task 1. We report mean and standard deviation of results. **Humans achieve considerably higher performance than models**.

Category	Def	Def + $q'(1)$ + $q(1)$	Def + $q'(k)$ + $q(k)$
Incomplete Information	88.00	72.00	76.00
Future Questions	53.00	62.00	65.00
Incorrect Information	71.00	66.00	67.00
Ambiguous	55.60	59.00	55.00
Unmeasurable	70.30	69.00	91.00
Avg	67.58	65.60	70.80

Table 3: Accuracy achieved by GPT-3 on Task 1 by using in-context examples.

Category	GPT-3
Incomplete Information	$98.00_{1.00}$
Future Questions	$92.50_{3.67}$
Incorrect Information	$53.50_{6.04}$
Ambiguous	$42.00_{8.42}$
Unmeasurable	$67.50_{4.47}$

Table 4: Accuracy achieved by GPT-3 on Task 2.

4.3 Models

We evaluate the performance of the following models: GPT-3 (Brown et al., 2020), Flan-T5 (Chung et al., 2022), and other models trained on NLI datasets. Since the performance of models varies with the task definition used for evaluation, we use 5 different definitions and report the mean and variance of the performance.

Definitions for Task 1: For Task 1, we use the following task definitions: (1) For the given question, identify if it has a definitive answer, (2) For the given question, output 'Yes' if it has a definitive answer otherwise output 'No'?, (3) Output 'Yes' if the given question has a definitive answer otherwise output 'No', (4) Identify whether you can give a definite answer for the given question, and (5) Can you give a definitive answer to the following question?

Definitions for Task 2: For Task 2, we use the following task definitions: (1) From the given two questions, identify which one has a definitive answer, (2) Which of the two below questions has a definitive answer?, (3) Which one of the two given questions has a definitive answer?, (4) You are given two questions, determine which one of the two questions has a definitive answer, and (5) Given the following two questions, find the one which has a definitive answer.

Definitions for Task 3: For Task 3, we use the following task definitions: (1) Justify why the following question does not have a definitive answer, (2) Explain why the following question fails to have a definitive answer, (3) Provide an explanation for the given question not having a definitive answer, (4) Why can't the following question be answered definitively?, and (5) Why there is no definitive answer to the following question?

We also explore the impact of providing in-context examples to the model. Specifically, we explore the following techniques:

- **Def**: Here, only the task definition is provided and the model needs to give its prediction for the test instances.
- **Def + $q'(1)$ + $q(1)$**: Here, a labeled (q', q) example pair is provided along with the task definition.
- **Def + $q'(k)$ + $q(k)$**: Here, k examples of q' and q are provided along with the task definition.

QnotA	GPT-3 Output	Ground Truth	Evaluation
Which countries participated in the fourth world war?	This question does not have a definitive answer **because there was no fourth world war**.	There was no fourth world war.	✓✓✓
When did Edmund Hillary climb the Gangkhar Puensum?	This question does not have a definitive answer because the mountain Edmund Hillary is believed to have climbed, Gangkhar Puensum, has never been officially summited and so **there is no definitive record of when he made the climb**.	Edmund Hillary never actually climbed the mountain Gangkhar Puensum.	✓✓✓
Look at the dog with one eye, does the dog have only one eye?	This question does not have a definitive answer **because it is impossible to know if the dog has had one eye since birth or has lost an eye since.**	There are two possible interpretations of the context: the dog can have one eye or you have to look at it from only one eye.	× × ×
Sam went for a walk with his friend in the red shirt. Who was wearing the red shirt?	This question does not have a definitive answer **because it does not provide any information about who else was on the walk or who else was wearing a red shirt.**	It is ambiguous to know whether Sam is wearing a red shirt or his friend.	× × ×
What technological advancements will be used at the Olympics in 2042?	This question does not have a definitive answer **because it is impossible to predict what technological advancements will be used at the Olympics in 2042**.	It is not possible to exactly predict the technological advancements that will be used in 2042.	✓✓✓

Table 5: Illustrative examples of justifications outputted by the model for randomly selected QnotA questions. The evaluation column corresponds to the evaluation of GPT-3 output with the ground truth justification performed by 3 authors.

4.4 Results and Analysis

4.4.1 Task 1

Table 2 shows the accuracy (mean and standard deviation) of various models on Task 1. It can be observed that humans perform notably well, especially on 'Incomplete Information', 'Unmeasurable', and 'Incorrect Information' categories. GPT-3 achieves considerably lower performance than human baseline.

Table 3 shows the impact of using in-context examples on the performance of GPT-3 model. From the results, it can be observed that variation in prompts indeed helps the model achieve better performance as compared to the baseline approach (i.e., **Def**). As observed in prior work (Mishra et al., 2022a; Wei et al., 2021; Parmar et al., 2022; Ouyang et al., 2022; Wang et al., 2022), adding examples to prompts helps in improving the performance. To this end, we experiment by adding examples and show that it (**Def** + $q'(k)$ + $q(k)$) leads to improvements on average. In particular, we observe notable performance improvements in the 'Future Questions', 'Incomplete Information', and 'Unmeasurable' categories. On average, the performance improves. Overall, Table 3 shows that adding in-context examples improves the performance and reduces the gap between model and human performance. However, there is a still considerable performance difference, implying the scope for future research in this direction.

Why is the Human Performance Low on Incorrect information category? We observe that the human performance is low on the incorrect information category. We attribute this to the humans' lack of information about topics which makes them unable to identify incorrect information in the question. For instance, a person who is unaware about basketball will say that the following question 'How many points did LeBron score for chicago bulls?' has a definitive answer. However, this question has incorrect information and thus doesn't have a definitive answer.

4.4.2 Task 2

Table 4 shows the accuracy achieved by GPT-3 on Task 2 where we provide two questions (QnotA and its corresponding QA instance) as input and the system is required to identify which one has a definitive answer. It shows that the GPT-3 is able to distinguish QnotA and QA instances of 'incom-

plete information' and 'future questions' categories really well. However, it struggles on 'ambiguous' and 'incorrect information' categories.

4.4.3 Task 3

Table 5 shows examples of justifications outputted by the GPT-3 model for Task 3 in which the system needs to provide a justification of why the given QnotA instance does not have a definitive answer. The evaluation column in the table corresponds to the evaluation of GPT-3 output (against the ground truth justification) performed by 3 authors. We find that in most cases (88% on average), the model is indeed able to generate the correct justifications for QnotA instances. Hence, despite not being able to accurately identify a QnotA question, on prompting GPT-3 to output a justification of why the given question doesn't have a definitive answer, it is able to provide a reasonable justification. Though in some cases the justification is not correct; for e.g. in the case of 'Sam went for a walk with his friend in the red shirt. Who was wearing the red shirt?', the output of GPT-3 is not exactly correct. Specifically, we find that the majority of the errors in generating justifications are on the 'ambiguous' and 'incorrect information' categories.

5 Scaling Up QnotA

Human-generated data is often expensive and time-consuming to collect. Furthermore, given the practical importance of data created in this work, we explore creating more such examples using automated means. To this end, we use the generation capabilities of GPT-3 to expand the dataset and provide in-context examples as shown in Figure 2. Specifically, the prompt for scaling up the dataset involves three components: topics, examples for each topic, and an instruction to generate new examples. In Table 6, we show examples of examples obtained using this approach. We (the authors) checked the validity of questions generated via this method and show the results in Table 7. Specifically, we randomly sampled 20 questions for each category and checked whether the question indeed doesn't have a definitive answer Table 7 shows that the majority of the synthetically created questions using the aforementioned method are valid. Thus, this method can be used to add more diversity to the dataset. This expanded dataset can be utilized for further explorations in this important area of research such as incorporating them

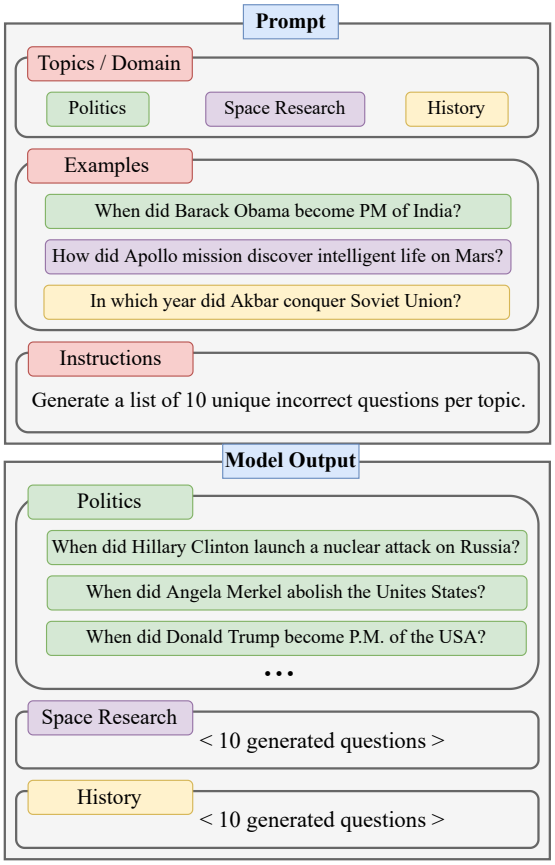

Figure 2: Schematic representation of our method for scaling-up QnotA.

in training the models. We note that the evaluation results reported in this paper are on the initial set of human-authored examples as they have been carefully created and comprehensively validated.

6 Conclusion

In real-world applications, users often ask questions that do not have definitive answers. Incorrectly answering such questions can have serious consequences and can thus hamper the system's reliability and trustworthiness. To study the ability of state-of-the-art NLP models to 'identify', 'distinguish', and 'provide a reasonable response' to such questions, we introduce QnotA, a dataset consisting of five different categories of questions that don't have definitive answers and formulate three different tasks. We showed that state-of-the-art models such as GPT-3 do not perform well on the task and achieve significantly lower performance than humans. Then, we demonstrated that this performance can be improved by providing in-context examples. We believe our work will encourage and facilitate future research in this important area and

Category	QnotA
Future Questions	What is the greatest invention of 2050?, What will be the biggest sporting event in 2044?, Who will be the most popular politician in 2040?
Incorrect Information	When did Hillary Clinton launch a nuclear attack on Russia?, When did Stephen Hawking discover a cure for cancer?
Ambiguous	She prepared the girl for the exam in June. When is the exam?, They stood watching the fireworks in the garden. Where were the fireworks?
Unmeasurable	What is the average density of the universe?, What would happen if humans could breathe in space?
Incomplete Information	Samira went to New York last weekend.What did Samira do in New York?, Jay proposed to Priya yesterday.Was it a surprise proposal?

Table 6: Illustrative examples of questions obtained using our scale-up techniques.

Category	Human Validation
Incomplete Information	20
Future Questions	20
Incorrect Information	20
Ambiguous	17
Unmeasurable	18

Table 7: Human validation performance on 20 randomly sampled questions generated using scale-up technique.

contribute towards the development of more robust NLP systems.

Limitations

Our dataset includes questions in only one language i.e. English. Furthermore, in our dataset, we have covered a diverse set of questions that don't have definitive answers but, it is in no way an exhaustive list. It can be further expanded with more categories of questions in future.

Ethical Considerations

The names used in this dataset are selected from the list of most common English names. In question creation, we ensure that all our contexts and questions describe realistic situations. Any bias observed in systems trained using our methods can be attributed to the source data. However, no particular sociopolitical bias is emphasized or reduced specifically by our data. No personal information from data creators has been collected during the creation of the dataset.

References

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Himanshu Gupta, Neeraj Varshney, Swaroop Mishra, Kuntal Kumar Pal, Saurabh Arjun Sawant, Kevin Scaria, Siddharth Goyal, and Chitta Baral. 2023. "john is 50 years old, can his son be 65?" evaluating NLP models' understanding of feasibility. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 407–417, Dubrovnik, Croatia. 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, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UNIFIEDQA: Crossing format boundaries with a single QA system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907, Online. Association for Computational Linguistics.

Najoung Kim, Phu Mon Htut, Samuel R Bowman, and Jackson Petty. 2022. $(qa)^2$: Question answering with questionable assumptions. *arXiv preprint arXiv:2212.10003*.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

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, pages 552–561. AAAI Press, Rome, Italy.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13480–13488.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022a. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022b. NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. In-BoXBART: Get instructions into biomedical multitask learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States. Association for Computational Linguistics.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

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.

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.

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.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019. QuaRTz: An open-domain dataset of qualitative relationship questions. 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 5941–5946, Hong Kong, China. Association for Computational Linguistics.

Neeraj Varshney and Chitta Baral. 2023. Post-abstention: Towards reliably re-attempting the abstained instances in qa. *arXiv preprint arXiv:2305.01812*.

Neeraj Varshney, Himanshu Gupta, Eric Robertson, Bing Liu, and Chitta Baral. 2023. A unified evaluation framework for novelty detection and accommodation in nlp with an instantiation in authorship attribution. *arXiv preprint arXiv:2305.05079*.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022a. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002, Dublin, Ireland. Association for Computational Linguistics.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022b. Towards improving selective prediction ability of NLP systems. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 221–226, Dublin, Ireland. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *EMNLP 2022*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

Albert Xu, Xiang Ren, and Robin Jia. 2022. Conal: Anticipating outliers with large language models. *arXiv preprint arXiv:2211.15718*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

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.