GOOAQ: Open Question Answering with Diverse Answer Types

Daniel Khashabi[1] **Amos Ng**
Tushar Khot[1] **Ashish Sabharwal**[1] **Hannaneh Hajishirzi**[1,2] **Chris Callison-Burch**[3]

[1]Allen Institute for AI, Seattle, WA, USA
[2]University of Washington, Seattle, WA, USA
[3]University of Pennsylvania, Philadelphia, PA, USA

Abstract

While day-to-day questions come with a variety of answer types, the current open question-answering (QA) literature represents isolated efforts on niche response types, with a heavy focus on specific kinds of short responses (people, places, etc.). To address this gap, we present GOOAQ, a large-scale dataset collected from Google questions and answers, containing 3 million questions with diverse answer types ranging from factual short answers to snippets to collections. Our human evaluation shows that 94% of the mined answers are accurate, enabling fine-tuning a pre-trained language model for answering GOOAQ questions. We use this dataset to study inherent differences between models producing different answer types, and observe interesting trends. For example, in line with recent work, LM's strong performance on GOOAQ's short-answer questions heavily benefits from annotated data. However, their surprisingly high quality in generating coherent and accurate answers for questions requiring long responses (such as 'how' and 'why' questions) is less reliant on observing annotated data and mainly supported by their pre-training. Moreover, we show that GOOAQ is a valuable training resource, resulting in strong performance on the recent ELI5 long-answers dataset. We release GOOAQ to facilitate further research on improving QA with diverse response types.[1]

1 Introduction

Research in "open" question answering (also referred to as open-response, open-domain, or direct answer QA) has resulted in numerous datasets and powerful models for answering questions without a specified context. This task requires the use of background knowledge either stored in the QA model or retrieved from large corpora or knowledge bases (Roberts et al., 2020; Lewis et al., 2021). Existing effort, however, involves isolated studies on niche answer types, mainly short responses and, in a few cases, long responses (Joshi et al., 2017; Lee et al., 2019; Bhakthavatsalam et al., 2021).

In contrast, many of the everyday questions that humans deal with and pose to search engines have a more diverse set of response types, as illustrated in Fig. 1. Their answer can be a multi-sentence description (a *snippet*) (e.g., 'what is' or 'can you' questions), a *collection* of items such as ingredients ('what are kinds of', 'things to') or of steps towards a goal such as unlocking a phone ('how to'), etc. Even when the answer is short, it can have rich types, e.g., unit conversion, time zone conversion, or a variety of knowledge look-up ('how much', 'when is', etc.).[2] Such answer type diversity is not represented in any existing dataset.

Motivated by this, we introduce GOOAQ (pronounced *guac* like *guacamole*), the first open QA benchmark containing *questions with all of the above answer types within a unified dataset, collected using the same, coherent process*. GOOAQ contains 3 million questions with short, snippet, or collection answers, such as the ones shown in Fig. 1. Besides supporting research on various types of answers, GOOAQ enables a quantitative study of the inherent differences in systems across different answer types.

GOOAQ questions are automatically mined from Google's search-autocomplete feature and thus, we hypothesize, represent popular queries of real-world interest. Such questions also trigger 'answer boxes' in the search results, containing responses deemed best by Google, which we extract and refer to as Google answers. Our human evaluation (§3.2) found the collected questions and answers to be of

[1]The dataset is available at https://github.com/allenai/gooaq under an appropriate license.

[2]In contrast, the short responses in existing datasets typically inquire about people, dates, and counts. For instance, 65% of Natural Questions (Kwiatkowski et al., 2019) begin with 'who', 'when', or 'how many'; cf. Fig 3.

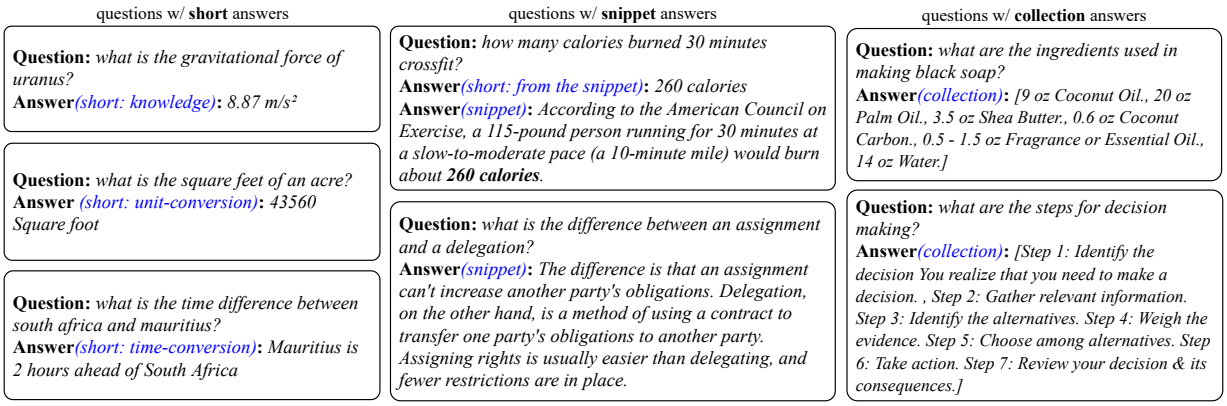

Figure 1: Examples from GOOAQ showing different types of the questions considered in this study. Each input is a natural language question, mapped to textual answer(s). The questions/answers come with answer type which are inferred from meta information of the search results.

high quality (over 94% valid answers).

GOOAQ provides a unified test bed to study inherent differences between questions. To do so, we fine-tune generative pre-trained language models (LMs) (Lewis et al., 2020; Raffel et al., 2020) on different subsets of GOOAQ, and ask whether models trained for different answer types:

(Q_1) *benefit similarly from pre-training?*
(Q_2) *benefit similarly from labeled data?*
(Q_3) *benefit similarly from larger models?*

To understand the contribution of pre-training, (Q_1), we train the powerful T5 language model (Raffel et al., 2020) on GOOAQ with a small amount of labeled data. While LMs struggle, as expected, in this setting on *short* response questions, they perform surprisingly well in generating *snippet* and *collection* responses.[3] We hypothesize this is because response fluency and coherence have a much higher weight in such questions, and these factors remarkably benefit from the LM pre-training objective. Regarding the value of labelled data, (Q_2), we observe the opposite trend: *short* response questions benefit consistently from increasing amounts of supervised (labeled) data, whereas both *snippet* and *collection* response questions show minimal gains (e.g., only 5-10% improvement when going from 2k training examples to 200k or even 2 million). Lastly, on the benefit of model size, (Q_3), we find larger models to be more effective in all cases as expected, but the gains are much more pronounced for *snippet* and *collection* response generation (20+%) as compared to *short* responses (5-10%), under human evaluation.

Additionally, we expect GOOAQ to facilitate further research on models for answering snippet and collection response questions. While the largest models we consider score surprisingly high on these questions, they are still far from reaching Google's quality under either automated or human evaluations. Importantly, due to little benefit observed from more labeled data on such questions, further progress requires rethinking the approach and devising new solutions.

Lastly, we find GOOAQ to be a valuable resource for training models. On the long-answer dataset ELI5 (Fan et al., 2019), T5 trained only on our snippet questions performs on par with state-of-the-art models trained on ELI5 data.

Our closing remarks describe why we aren't simply replicating an existing QA system at Google, place our findings in context, and discuss future uses of GOOAQ, such as creating a neural knowledge-base or a question generation system.

Contributions. Our contributions are threefold:

1. We present GOOAQ, a collection of 3 million question-answer pairs with a diverse set of answers, along with a crowdsourced assessment of its quality.
2. We benchmark state-of-the-art models on GOOAQ, both in terms of automatic and human judgments, and observe remarkable differences in how models behave on different answer types.
3. We demonstrate that GOOAQ is also a valuable model training resource by showing strong generalization to ELI5 (Fan et al., 2019).

[3] Over 30-40% of our best model's snippet and collection answers were preferred by crowdworkers over Google's answers; achieving 50% here would mean parity with Google.

2 Related Work

A closely related work is the Natural-Questions (NQ) dataset (Kwiatkowski et al., 2019; Lee et al., 2019) which contains questions written by Google users, and answers that were manually extracted from Wikipedia articles. While our questions (extracted via autocomplete) were also likely frequently asked by Google users, our dataset represents a different and wider distribution of questions (§3.3), likely because it encompasses different classes of answers, particularly snippet and collection responses. Specifically, while NQ is dominated by 'who', 'when', and 'how many' questions (cf. Fig. 3(d)), GooAQ has notably few 'who' questions and a substantial portion of questions starting with 'how to', 'what is', 'what does', 'can you'.

One notable QA dataset with long-form responses is ELI5 (Fan et al., 2019; Krishna et al., 2021), containing questions/answers mined from Reddit forums. In contrast, GooAQ is collected differently and is several orders of magnitude larger than ELI5. Empirically, we show that models trained on GooAQ transfer surprisingly well to ELI5 (§5.3), indicating GooAQ's broad coverage.

It is worth highlighting that there is precedent for using search engines to create resources for the analysis of AI systems. Search engines harness colossal amounts of click information to help them effectively map input queries to a massive collection of information available in their index (Brin and Page, 1998; Joachims, 2002; Berant et al., 2013; Joachims et al., 2017). Although academic researchers do not have direct access to information collected from the users of search engines, search results can act as a proxy for them and all the complex engineering behind them. In particular, the GooAQ dataset used in this study probably is *not* representative of *a single* QA system in Google; on the contrary, we hypothesize, this data is produced by a complex combination of many systems, various forms of user feedback, as well as expert annotation/verification of highly popular responses.

3 GooAQ dataset

We describe how GooAQ was collected, followed by dataset statistics and quality assessment.

3.1 Dataset Construction

Constructing this dataset involved two main steps, extracting questions from search auto-complete and extracting answers from answer boxes.

3.1.1 Query Extraction

To extract a rich yet natural set of questions we use Google auto-completion.[4] A similar strategy was also used by Berant et al. (2013), albeit in the context of a slightly different study. We start with a seed set of question terms (e.g., 'who', 'where', etc.; the complete list is in Appendix A.) We bootstrap based on this set, by repeatedly querying prefixes of previously extracted questions, in order to discover longer and richer sets of questions. Such questions extracted from the autocomplete algorithm reflect popular questions posed by users of Google. We filter out any questions shorter than 5 tokens as they are often incomplete questions. This process yields over ∼5M questions, which were collected over a span of 6 months. The average length of the questions is about 8 tokens.

3.1.2 Answer Extraction

To mine answers to our collected questions, we extract the Google answer boxes shown on top of the search results when the questions are issued to Google. There are a variety of answer boxes. The most common kind involves highlighted sentences (extracted from various websites) that contain the answer to a given question. These form the *snippet* and *collection* answers in GooAQ. In some cases, the answer box shows the answer directly, possibly in addition to the textual snippet. Similarly, *unit-conversion* and *time-conversion* they each have distinct answer boxes. Some technical details of the answer extraction is included in Appendix B.

After the answer extraction step, we have all the necessary information to create a question in GooAQ, such as the examples in Fig. 1.

Answer Type Categories. We use the HTML tags of the search results to infer answer type tags for each answer. The overall list of types are shown in Table 1 (examples in Fig. 1). We define 'short' response questions to be the union of 'knowledge', 'unit-conversion', 'time-conversion', and short answers from the 'snippet' responses.

Table 1 summarizes various statistics about GooAQ broken down into different question/answer types. Of the 5M collected questions, about half resulted in successful answer extraction from answer boxes. The largest type of questions received 'snippet' answers with over 2.7M responses (examples shown in the left-most column of Fig. 1). The other major category is 'collection'

[4] http://google.com/complete/search?client=chrome&q=...

answers with 329k questions (examples shown on the right-most column of Fig. 1).

Answer types	Count	% of valid questions	% of valid answers
short answers	275k	-	-
↳ unit conversion	45k	96.9	90.9
↳ time conversion	2.5k	93.2	70.9
↳ knowledge	32k	96.3	84.1
↳ snippet (short)	196k	98.4	76.0
snippet	2.7M	98.5	95.5
collection	329k	99.7	98.9
Overall	3.1M	98.6	94.5

Table 1: Statistics of different answer types in GOOAQ (§3.3) and their quality evaluation by crowdworkers (§3.2). According to human ratings, a very small percentage of the questions are invalid (first column). Among the valid questions, a substantial portion are deemed to have valid answers.

3.2 Quality Assessment of GOOAQ

We perform a crowdsourcing experiment to assess the quality of the extracted questions and their answers. We use Amazon Mechanical Turk (AMT) to annotate about 2.5k randomly selected question-answer pairs. The annotators were asked to annotate (1) whether a given question makes sense and, if so, (2) whether the provided answer is complete.

Annotation details. Since our task is focused on English, we required workers to be based in a country with a population predominantly of native English speakers (e.g., USA, Canada, UK, and Australia) and have completed at least 5000 HITs with $\geq$ 99% assignment approval rate. Additionally, we have a qualification test with half-a-dozen questions all of which need to be answered correctly by our annotators. To prevent biased judgements, we also ask annotators to avoid using Google search (which is what we used to mine GOOAQ) when annotating the quality of shown instances. Each example is annotated by 3 independent annotators and aggregated via a majority vote of the 3 labels.

Assessment results. We compute aggregate statistics for (1) average rating of questions and (2) average rating of the answer quality, among valid questions. As can be seen in the results in Table 1 only a small percentage of the questions were deemed 'invalid'. Additionally, among the 'valid' questions, a high percentage of the answers were deemed high-quality for most of the question/answer types. This indicates a reasonable quality of GOOAQ question-answer pairs, as evaluated directly, independent from any systems. (Examples of invalid questions/answers are provided in Appendix C.)

3.3 Dataset Analysis

To better understand the content of GOOAQ, we present several distributions from the data. Fig. 2 shows the length distribution of GOOAQ questions and that of NQ (Kwiatkowski et al., 2019). While a vast majority of NQ questions contain 8-10 tokens, GOOAQ questions have a somewhat broader range of lengths.

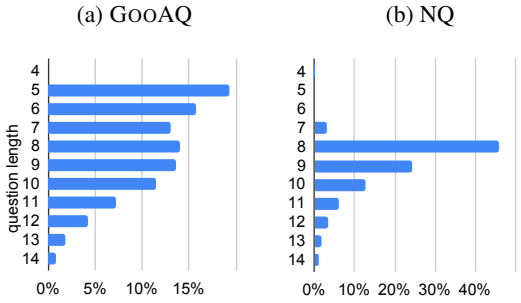

Figure 2: Comparison of question length distributions

To gain insights about the type of questions, we study the distribution of the most frequent opening bigrams of the questions (Fig. 3). Among the *short* answer questions, the majority are information-seeking questions about counts ('how many'), places ('where is'), values ('how much'), and people ('who is'). They also include 'what is' questions, which can cover a wide variety of open-ended queries with short answers (e.g., *what is the time difference . . .?*, *what is the length of X?*, etc.). Among the *snippet* questions, the dominant pattern is 'what is', which typically is an open-ended question about entities (e.g., *'what is X?'* or *'what is the difference between X and Y?'*). Among the *collection* response questions, most questions are about steps or ingredients needed to accomplish a goal ('how to' and 'what are'). A comparison with the bigram distribution of NQ (Fig. 3; right) highlights that GOOAQ represents a different and wider class of questions. Specifically, NQ has many 'who', 'when', and 'how many' questions, while GOOAQ dominantly contains 'how' and 'what' questions, which typically require explanatory responses.

In terms of the different reasoning types, GOOAQ has an extremely long-tail of reasoning challenges, due to our data collection procedure. For example, we observed many challenges such as

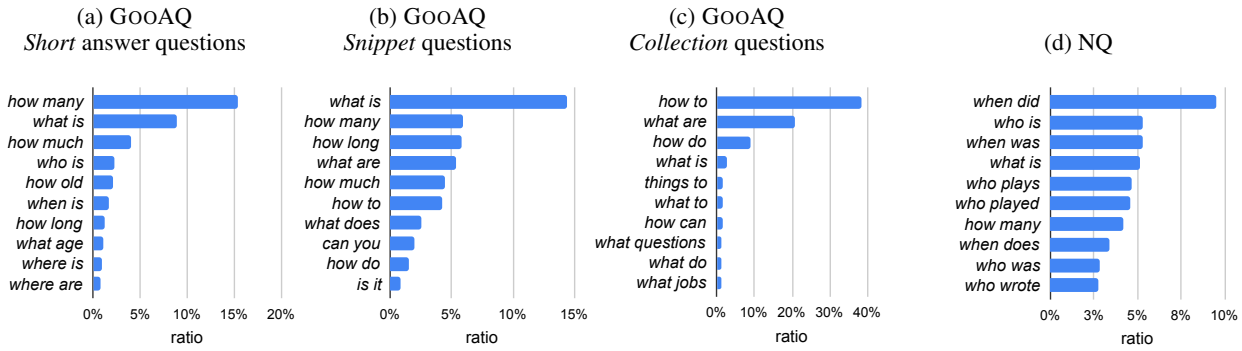

Figure 3: The distribution of common bigrams in questions of GooAQ (a,b,c) vs. NQ (d).

application of mathematical definitions (Q: *'what is the multiplicative inverse of 10?'* A: *'1/10'*), linguistic definitions (Q: *'a man who looks after cattle?'* A: *'cowherd'*; Q: *'a man who protects sheep?'* A: *'Shepherd'*), comparisons (Q: *'are boiling and evaporation the same?'*; Q: *'what is the difference between night sky and day sky?'*), instantiation (Q: *'what is an example of kinetic energy?'*), etc., to name a few. Because of the long tail of reasoning phenomena, a detailed analysis would require careful human annotations which we leave for future work.

4 Task Setup and Models

GooAQ naturally forms a dataset for the task of open QA, where the input is a question and the output is its answer. Unlike the reading comprehension setting, the context for answering the question is not provided as part of the input. In particular, we consider the so-called 'closed-book' setup (Roberts et al., 2020) where the model (e.g., a language model) is expected to use background knowledge stored within it, without access to any additional explicit information retrieval mechanism.[5]

4.1 Problem Setup

We split GooAQ into three sub-tasks: ($\mathcal{T}_{short}$) *short* responses questions, ($\mathcal{T}_{snippet}$) *snippet* responses questions, and ($\mathcal{T}_{collection}$) *collection* response questions. We train and evaluate models for each of these sub-tasks separately. We define them as different sub-tasks since by merely reading the questions it might not be clear whether its response should be short, a snippet, or a collection,

Data splits. For each sub-task, we randomly sample *test* and *dev* sets such that each evaluation split contains at least 500 instances of each response type. We experiment with varying training data sizes to better understand the value of labeled data. Lewis et al. (2021) have shown that leakage from training data to the evaluation sets often results in unrealistically high scores. To minimize this issue, we create training splits by selecting the most *dissimilar* instances to our evaluation splits. The measure of *similarity* for each training instance is computed as the maximum amount of token-overlap with any of the instances in the test/dev set (computed over both questions and answers). Using the most *dissimilar* subset of the training instances, we create training splits of the following sizes: 2k, 20k, 200k. For $\mathcal{T}_{snippet}$, we also have a 2M training set since this sub-task has more data.

4.2 Evaluation Metrics

Automatic evaluation. We use the ROUGE-L metric (Lin, 2004) , which is a common metric for assessing the quality of models for text generation tasks. The results of the automatic evaluation for each sub-task are shown in the top row of Fig. 4.

Human evaluation. We additionally perform human evaluation which is generally known to provide more accurate evaluation for generated text. Specifically, we ask crowdworkers to indicate if they prefer the predicted answer by the model or the Google answer for each question (without revealing the source of the answers).

The annotation interface is shown in Fig. 5, which is essentially the same template used for the quality assessment of the dataset (§3.2), except that here the crowdworkers are shown a *pair* of

[5] In our early experiments, we considered information-retrieval (IR) systems in conjunction to LMs (i.e., an 'open-book' setup). We observed that IR results are quite noisy for most open questions. Hence, a system trained with the retrieved documents did not benefit from them (the model learned to ignore the noisy retrieval results). Similar observations were also made by Krishna et al. (2021, Sec3.1) ("generations are similar irrespective of type of retrievals").

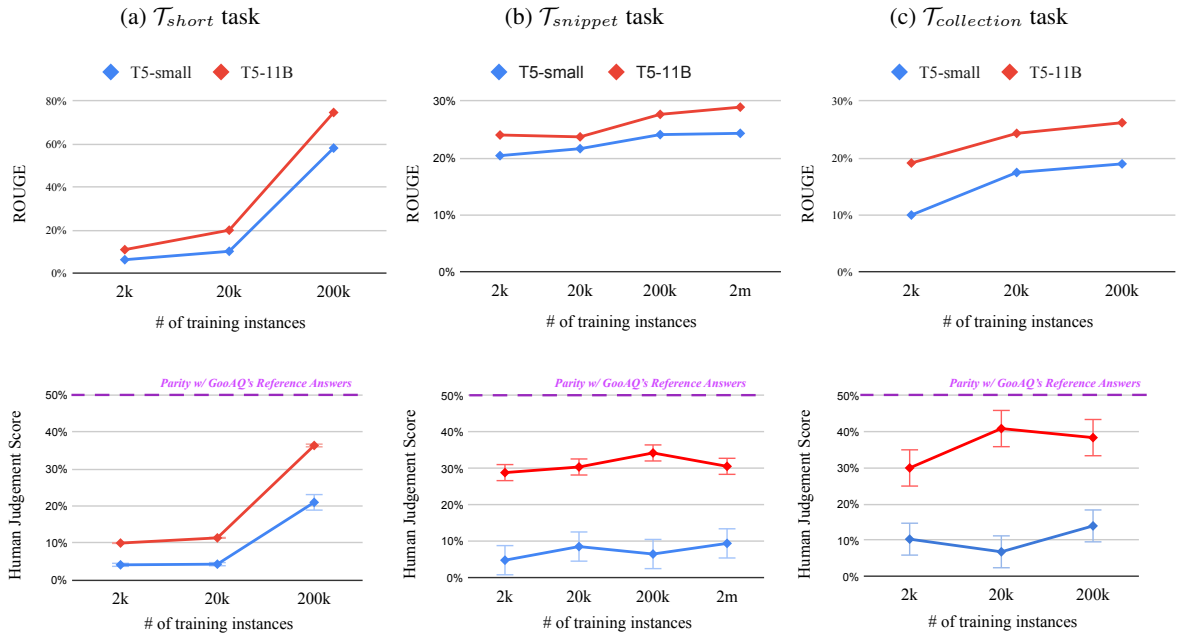

Figure 4: Evaluation of T5 (small,11B) models on different sub-tasks of GooAQ via *automatic* metrics (top) and *human* judgements (bottom). For human evaluation, 50% is the border at which the model output and the ground truth responses are indistinguishable. The short-answer sub-tasks ($\mathcal{T}_{short}$; left) have a relatively low performance when supervised with $2k$ instances. However, they benefit more than the long-answer sub-tasks ($\mathcal{T}_{snippet}$ & $\mathcal{T}_{collection}$) from more labeled data. Additionally, we observe that the gap between the two systems is bigger in terms of human evaluation (compared to the corresponding gap in terms of automatic evaluation), especially in the *long* response tasks (middle & right).

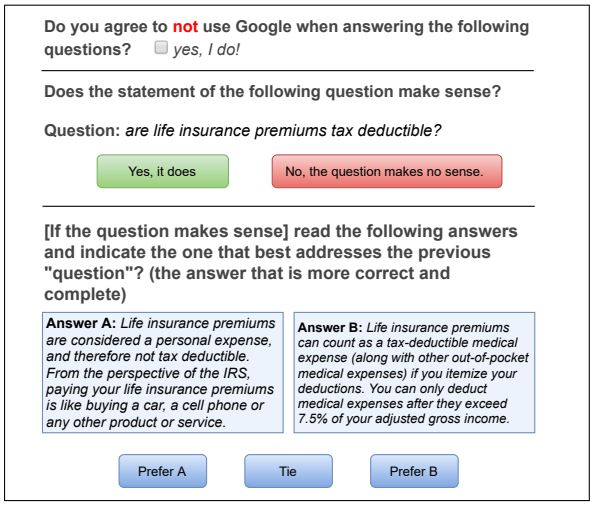

Figure 5: Crowdsourcing interface used for human assessment of our baselines (§4). We use a similar template (with a single answer) to estimate the quality of GooAQ (§3).

responses for each question—the reference answer (extracted from Google) and the one generated by the model—turning the task into a *comparative* one. Before annotating each instance, we remind the annotators to avoid using Google. Then we ask them to check if the provided question is clear enough and makes sense. Upon indicating 'yes', they choose between the Google answer, the generated answer by our model, or indicate that they are equally good (by selecting 'tie').

For each question, we obtain annotations from 5 independent annotators and aggregate via a majority vote.[6] The model receives a credit of 1 if the majority vote favors the model's prediction, 0.5 if the majority vote is the 'tie' label, and 0 otherwise. The overall accuracy score for the model is computed by averaging instance-level scores, after discarding questions annotated as invalid ('this question makes no sense').

The resulting *human-evaluation* metric indicates how often were model predictions preferred over Google's answers. In this evaluation, 50% is the mark where the annotators are not able to distinguish the model's responses from Google's answers in any meaningful way. The results of human evaluation are shown in the bottom row of Fig. 4.

[6]Ties occurred infrequently (e.g., in 6% of the cases when evaluating our largest T5 model) and were broken at random.

4.3 Models

For our evaluation, we use the T5 model (Raffel et al., 2020), a recent text-to-text framework that has achieved state-of-the-art results on a variety of tasks, including open QA (Roberts et al., 2020). The models are trained to produce answer string, given the question string as input. We use two model sizes that capture the two extremes: the smallest model ('small') and the largest model ('11B'). Both models were trained for 20k steps on the training splits, dumping checkpoints every 2k steps (with 196,608 tokens per batch on v3-128 TPUs) with the default hyperparameters. We select the checkpoint with the highest score on the 'dev' set and report its corresponding 'test' score.

5 Empirical Results and Analyses

In this section, we evaluate the behavior of models for various answer types (§5.1). We further show how GOOAQ can support research in answering questions with long answers (§5.2;§5.3).

5.1 Models vs. Various Answer Types

(Q_1) **Model pre-training is surprisingly effective on the snippet and collection answer subtasks** Both automatic and human evaluations of these two classes of questions (Fig. 4; middle & right) demonstrate that the T5-11B model is surprisingly effective at answering them, with only 2k training examples. For example, crowdworkers even prefer the model's answer over Google's in 30% of the cases.[7] This is in contrast with short answer questions, where the model's accuracy is only around 10% and crowdworkers prefer Google's answers in about 90% of the cases.

To understand this observation, one needs to put into perspective several factors that are at play. First, short answer questions typically ask for encyclopedic knowledge and, therefore, *correctness* of the answers matters the most. In snippet and collection questions, we suspect *coherence* of the response carries a heavier weight. This is partly due to the nature of the questions, which can be responded to in a variety of ways. For example, the snippet response to the question of *how many calories burned 30 minutes crossfit?* (Fig. 1) could refer to a range of calorie consumption, depend on the choice of activity during crossfit, or vary by the attributes of the person working out. All of these responses would be equally correct.

(Q_2) **Labeled data is more helpful for short answer questions.** Based again on both the automatic and human evaluations (Fig. 4; left), the performance of both small and 11B parameter models on the short response questions quickly improves as we increase the amount of training data, especially beyond 20k. This is in contrast with snippet and collection questions, where even 200k labeled instances don't appear to help much, indicating that in these question types, model pre-training contributes more than labeled data does.

(Q_3) **Human evaluation accentuates the gap between the 'small' and '11B' models, especially on snippet and collection response questions.** This is visually evident from the gap between the blue and red curves in the bottom row vs. the top row of Fig. 4. This is compatible with recent work of Min et al. (2021), who also observed that the gap between two reasonably different systems is bigger when using human evaluation. We hypothesize this is due to the crudeness of automatic evaluation metrics, and an indication of the necessity of human evaluation to distinguish between nuanced differences among generated responses.

What is perhaps more interesting (and not evident from prior work) is that the gap between automatic and human evaluation is larger for the snippet and collection questions than short answer questions, especially for the T5-small model. This is, at least partly, due to the inaccuracy of automatic metrics in evaluating long text.

5.2 GOOAQ as a challenge for LMs

One can view GOOAQ as a challenge for NLP, for building self-contained models that achieve performance comparable to Google's answers.

As mentioned earlier, our human evaluation measures the comparative quality of the model predictions and our reference responses (Google's answers). Hence, a value of 50% in this evaluation is an indication that the predictions are on par with (i.e., indistinguishable from) the ground-truth responses (defined in 'human-evaluation' §4.2).

As the bottom row of Fig. 4 shows, the T5-11B model comes quite close to Google's answers but is still not quite at par with it. We hope this gap will encourage further research in building stronger models, especially for the snippet and collection

[7] Across all experiments, the model's and Google's answers were deemed a "tie" in fewer than 10% of the cases.

answer questions where more labeled data doesn't appear to be a promising way to increase accuracy.

5.2.1 Error Analysis

To gain an intuition about the mistakes made by the models, we conducted a small-scale errors analysis of model predictions. For each model, we (one of the authors) annotated 30 predictions, and labeled them with the following error categories inspired from existing evaluations of text summarization (Chaganty et al., 2018): *incompleteness*, indicating the lack of expected substance in the prediction; *redundancy*, indicating repeated content; *hallucination*, indicating existence of made-up statements; and *incoherence* indicating the existence of grammatical errors (examples in Appendix D).

Model	Incompleteness	Redundancy	Hallucination	Incoherence
T5-small	52.5	65.0	47.5	2.5
T5-11B	22.5	8.3	18.3	0.0

Table 2: Error distribution for the two models

The results of our error analysis are summarized in Table 2. As expected, the 'small' model makes more errors across all categories, and suffers particularly from *redundancy* and *incompleteness*. Overall, both models have very little *incoherence*, which is to be expected from their strong pre-training.

5.3 Extrinsic Utility of GooAQ

To showcase the value of GooAQ as a model training resource, we train our models on questions from GooAQ and evaluate them on ELI5 (Fan et al., 2019), a relatively recent dataset with long-answer questions extracted from Reddit posts.

Model	Supervision	Uses IR?	Score
T5-small	GooAQ (no ELI5)	no	21.7
T5-11B	GooAQ (no ELI5)	no	22.9
T5-small	ELI5	no	19.0
T5-11B	ELI5	no	22.7
RAG*	ELI5	yes	14.1
RT+REALM*	ELI5	yes	23.4

Table 3: Evaluation of our models on ELI5 dataset. Results indicated with * are reported from prior work (Krishna et al., 2021). T5 fine-tuned on GooAQ performs well on ELI5, another long-answer dataset.

Our evaluation, summarized in Table 3, shows that both our small and 11B T5 models trained on GooAQ's snippet-answer subset (no training on ELI5) perform quite well (21.8 and 22.9, respectively) when evaluated on ELI5. They are even *better than* the same architectures trained with ELI5's own training data (19.0 and 22.7, resp.) and on par with retrieval based state-of-the-art models (23.4). Complementary to these results, a T5-11B model trained on ELI5 and evaluated on GooAQ results in 22.6%, much lower than ∼28.9% in Table 4.

We hypothesize that despite GooAQ being collected differently than ELI5, a notable portion of ELI5 is covered by GooAQ, indicating good coverage of common questions posed by ordinary users.

6 Closing Remarks

We studied open QA under diverse response types. To this end, we collected GooAQ, a very large set of QA pairs mined from Google, with a variety of short and long answer types, all of which are collected using a unified, coherent process, enabling a cross-type comparison. The auto-complete system used for our question collection likely reflects a natural distribution of questions asked by users.

We benchmarked two variants of a state-of-the-art self-contained text generation model (T5, without retrieval) on three different sub-tasks of GooAQ: short, snippet, and collection response questions. Our analysis, using both automatic and human evaluations, brings out the distinct behavior of LMs on long and short response questions. For example, while short response models benefit heavily from more labeled data, the surprisingly strong performance of long response models is driven mostly by their pre-training. We also demonstrate that GooAQ is a valuable resource for training models by showing high performance on an extrinsic task, ELI5, while using only GooAQ data for training.

Scope of our conclusions. One must be careful in taking our specific conclusions out of the context of this study (i.e., the dataset at hand, the models, the evaluation metrics used, etc.). While we expect our findings to be fairly general, it may be possible to come up with a different long-form QA dataset where the trends across answer types differ.

Knowledge leakage across train and evaluation sets has been shown to significantly inflate performance numbers on recent open QA datasets (Lewis et al., 2021; Emami et al., 2020). Similar concerns

have motivated our careful training/evaluation splits of the data (§4) and experiments with varying training set sizes. Nevertheless, we found it challenging to define (and assess) the amount of such leakage, and welcome such studies on GOOAQ.

Are we mimicking Google's QA? A reader might question the value of this work by noting that the website from which GOOAQ was mined had likely also used a QA system to begin with. In other words, are we basically reverse-engineer Google's internal QA system (Kilgarriff, 2007)?

While we (the authors) are not aware of how Google answer box system works, we suspect that it is much more complex than a single QA system built using a single LM like T5 (Raffel et al., 2020). The system, besides incorporating one or more QA models, likely makes heavy use of implicit user feedback (e.g., information contained in billions of clicks, the structure of web links, etc.), in addition to explicit feedback from users and possibly some expert curation of answers to common questions. Moreover, Google's system may decide which questions to display answers for, and probably limits itself to the answers that it is most confident in.

Thus, the data in Google's answer boxes likely captures a variety of signals that contribute towards its high-quality. We believe aiming for a 'standard' NLP QA system that's on par with Google QA is therefore a challenging and worthwhile goal.

Future uses of GOOAQ. One challenge in the progress on long-form QA is response evaluation. To facilitate future work on GOOAQ and replicability of our human evaluations, we have released the templates used for crowdsourcing human judgements. Efforts on text generation tasks such as ours will benefit from—and should in turn benefit advances in—proposals for streamlining human evaluation of models (Khashabi et al., 2021).

We hope our analysis and data will benefit the understanding of and further development of QA systems for dealing with diverse response types.

While we used GOOAQ for the purposes of QA, we expect this data to have a variety of use-cases, such as building a *knowledge-base* accessible via question queries (Bosselut et al., 2019), creating a better question generation system, etc. We leave such investigation to future work.

Acknowledgement

The authors would like to thank Sihao Chen, Peter Clark, and Oyvind Tafjord for their help throughout this project. TPU machines for conducting experiments were provided by Google.

Chris Callison-Burch was supported in part by the DARPA KAIROS Program (contract FA8750-19-2-1004), the DARPA LwLL Program (contract FA8750-19-2-0201), and the IARPA BETTER Program (contract 2019-19051600004). Hanna Hajishirzi was supported in part by ONR N00014-18-1-2826, Allen Distinguished Investigator Award, and NSF CAREER award. Approved for Public Release, Distribution Unlimited. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, IARPA, or the U.S. Government.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of EMNLP*, pages 1533–1544.

Sumithra Bhakthavatsalam, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, and Peter Clark. 2021. Think you have Solved Direct-Answer Question Answering? Try ARC-DA, the Direct-Answer AI2 Reasoning Challenge. *arXiv preprint arXiv:2102.03315*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of ACL*, pages 4762–4779.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30:107–117.

Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of ACL*, pages 643–653.

Ali Emami, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. An analysis of dataset overlap on winograd-style tasks. In *Proceedings of COLING*, pages 5855–5865.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of ACL*, pages 3558–3567.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of KDD*, pages 133–142.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, pages 4–11. Acm New York, NY, USA.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of ACL*, pages 1601–1611.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A Smith, and Daniel S Weld. 2021. GENIE: A leaderboard for human-in-the-loop evaluation of text generation. *arXiv preprint arXiv:2101.06561*.

Adam Kilgarriff. 2007. Googleology is bad science. *Computational linguistics*, 33(1):147–151.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of NAACL*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *TACL*, 7:453–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of ACL*, pages 6086–6096.

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 ACL*, pages 7871–7880.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of EACL*, pages 1000–1008.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, et al. 2021. NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned. *arXiv preprint arXiv:2101.00133*.

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. *JMLR*, 21(140):1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *Proceedings of EMNLP*.

A Query Terms

The list of terms used for bootstrapping questions: “who”, “whom”, “whose”, “what”, “which”, “when”, “where”, “why”, “how”, “should”, “would”, “wouldn’t”, “can”, “can’t”, “will”, “won’t”, “aren’t”, “do”, “does”, “has”, “have”, “am”, “are”, “is”, “shouldn’t”, “isn’t”, “could”, “couldn’t”, “does”, “don’t”, “must”, “may”, “ought”.

B Extracting Answers from Google

For technical reasons, the answer extraction was done in two steps. (1) We first scrape the search results for all of our questions. This is the main extraction bottleneck as there is no official APIs to provide the answer boxes. Therefore, one needs to extract them directly from the HTML search results. We use Selenium[8] which simulates browser experience. Note one cannot send too many queries to Google in a short span of time (due to various query limits). Therefore, we ensured to have enough delays between our queries (otherwise, we’d be blocked). Overall, this extraction process was done in 3 months. Subsequent to extracting the search HTML results, (2) we extract answer strings from the HTML content of the search results. Answer types are also inferred at this stage, based on the HTML tags around the answer.

C Invalid Questions and Answers

Based on the human evaluation of GooAQ in §3.2, we should example of erroneous instances. Figure 6 shows examples of invalid questions. Often the questions are deemed invalid since they’re under-defined or significantly deviate from the proper English. Figure 7 shows examples of invalid answers (to valid questions). Invalid answers often do not sufficiently address the topic of the given question.

Type: curr-conv
Question: 1 euro is hoeveel nok?
Question: how much is 45 in nigeria money?

Type: time-conv
Question: 2 am eastern standard time?
Question: what is the difference between china and republic of china?

Type: knowledge
Question: what age is beauty and the beast?
Question: acdc who made who live at donington?

Type: snippet
Question: have mercy on me o god according to your loving kindness?
Question: dating a guy who is selfish?

Type: collection
Question: what are some areas of improvement?
Question: can sıkıntısına ne iyi gelir?

Figure 6: examples of invalid questions

Type: time-conversion
Question: what is the difference between mexican and spanish?
Answer: Madrid, Spain is 7 hours ahead of Mexico City, CDMX, Mexico

---

Type: unit-conversion
Question: what is 12 pm in spanish?
Answer: 13:00 Saturday, in Madrid, Spain

Type: snippet (short)
Question: how many working days in january 2020 malaysia?
Answer: 262 working days

Type: knowledge
Question: aids and hiv are acronyms for?
Answer: HIV/AIDS

Type: snippet
Question: are ralph lauren jackets good?
Answer: Connoisseur. They are made by Corneliani in half-canvas construction. If you like them, they fit and you can afford them, they are a great value at 75% off. I think they are still good at 50% off.

Type: collection
Question: how long does it take to get through john wayne airport?
Answer: [’0-10 Min.’, ’11-20 Min.’, ’21+ Min.’]

Figure 7: examples of invalid answers (to valid questions)

[8] https://github.com/SeleniumHQ/selenium/

D Error Analysis of Model Predictions

Here are examples of *hallucination*, *redundancy*, and *incompleteness* errors in model predictions based on our human analysis discussed in §5.2.1. As noted earlier, all models exhibit very few *incoherence* errors, mainly because this aspect is directly addressed during model pre-training.

Question: how much does it cost to get court married?
Google ans: Your application must be notarized. Please include at least one phone number where you can be reached during the day. Enclose a check or money order for $69 payable to King County. The check or money order must include your address.
Predicted ans: The cost of a court marriage is between $40,000 and $50,000. The cost of a court marriage is between $40,000 and $50,000. The cost of a court marriage is between $40,000 and $50,000.
Explanation: In this example, the model has made up numbers that do not seem reasonable (hence, a 'hallucination' error.) Additionally, it is repeating the same point (hence, a 'redundancy' error.)

Question: who went to the superbowl last year?
Google ans: Here's your answer: The Eagles won Super Bowl LII on Feb. 4, 2018, beating the Patriots 4133 to win the league championship.
Predicted ans: The Super Bowl was a national championship game between the New York Giants and the New York Giants. The team was a member of the New York Giants, who starred in the Super Bowl in the 1970s.
Explanation: In this example, there is repetition in the model's prediction (hence, a 'redundancy' error.)

Question: what courses do you need to be a kinesiologist?
Google ans: To become a kinesiologist, you typically need either a bachelor's or a master's degree in Kinesiology or a related field, such as Human Kinetics, Exercise Science, or Exercise Physiology.
Predicted ans: To become a kinesiologist, you will need to complete a four-year degree in kinesiology, biology, chemistry, chemistry, physics.
Explanation: In this example, the model provides some relevant points, but the response is not as detailed as the gold response (hence, an 'incompleteness' error.)

E Fine-grained Evaluation of Short-Answer Questions

Fig. 8 shows the breakdown of the automatic evaluation for different types of short response questions. As expected, certain niche question types (such as 'unit-conversion') benefit the most from labeled data. In contrast, open-ended question types (such as 'knowledge' lookup) benefit less from more labeled data.

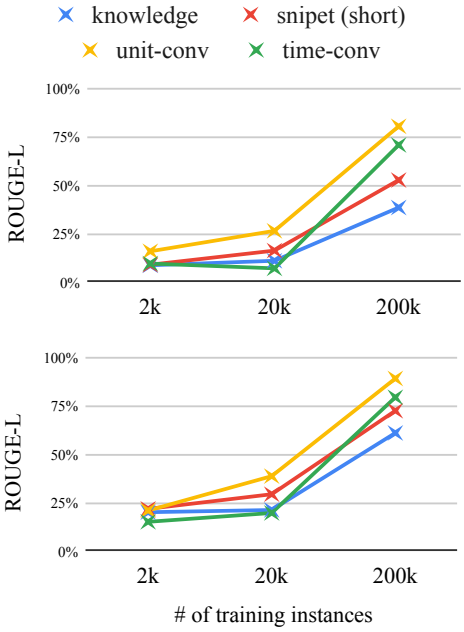

Figure 8: *Automatic* evaluation of T5 (small: top, 11B: bottom) models on different types of the questions included in short-answer sub-task ($\mathcal{T}_{short}$). 'unit-conversion' questions benefit the most from more labeled data, while 'knowledge' lookup questions are the opposite.

Question: *how do you change your background in a zoom meeting?*
Answer*(snippet)*: *While in a Zoom meeting, tap More in the controls. Tap Virtual Background. Tap the background you would like to apply or tap + to upload a new image. The background will be automatically applied.*

Question: *what is an assignment what is the difference between an assignment and a delegation?*
Answer*(snippet)*: *The difference between assignment and delegation is that an assignment can't increase another party's obligations. Delegation, on the other hand, is a method of using a contract to transfer one party's obligations to another party. Assigning rights is usually easier than delegating, and fewer restrictions are in place.*

Question: *what happens if a person dies without a will?*
Answer*(snippet)*: *A person who dies without a will is known as 'dying intestate'. ... Sorting out an estate without a will usually takes more time. So, the sooner you apply for probate, the sooner the you can distribute the estate to heirs. If there are no surviving relatives, the person's estate passes to the Crown.*

Question: *what is the difference between map and chart?*
Answer*(snippet)*: *A map usually represents topographical information. A chart is used by mariners to plot courses through open bodies of water as well as in highly trafficked areas. ... A map, on the other hand, is a reference guide showing predetermined routes like roads and highways.*

Question: *does drinking a lot of water flush out calories?*
Answer*(snippet)*: *Some research indicates that drinking water can help to burn calories. In a 2014 study, 12 people who drank 500 mL of cold and room temperature water experienced an increase in energy expenditure. They burned between 2 and 3 percent more calories than usual in the 90 minutes after drinking the water.*

Question: *if it's 10 am cst what time is it est?*
Answer*(short:time-conversion)*: *11:00 AM Wednesday, Eastern Time (ET)*

Question: *what is the difference between australia and america?*
Answer*(short:time-conversion)*: *Canberra ACT, Australia is 14 hours ahead of Washington, DC*

Question: *10 am central to mst?*
Answer*(short:time-conversion)*: *9:00 AM Thursday, Mountain Time (MT)*

Question: *what is the difference between bangalore and mangalore?*
Answer*(short:time-conversion)*: *here is no time difference between Bengaluru, Karnataka, India and Mangalore, Karnataka, India*

Question: *how high is 1.8 meters in inches?*
Answer*(short:unit-conversion)*: *70.8661 Inch*

Question: *how many cc's are there in a liter?*
Answer*(short:unit-conversion)*: *1000 Cubic centimeter*

Question: *how long is 1.6 cm in mm?*
Answer*(short:unit-conversion)*: *16 Millimeter*

Question: *how many centimeters are there in 1 kilometre?*
Answer*(short:unit-conversion)*: *100000 Centimeter*

Question: *how high is the great smoky mountains?*
Answer*(short: knowledge)*: *6,644'*

Question: how long can a cat be pregnant for?
Answer*(short: knowledge)*: *58 – 67 days*

Question: *are koala bears an endangered species?*
Answer*(short: knowledge)*: *Not extinct*

Question: *chevy is from what country?*
Answer*(short: knowledge)*: *Detroit, Michigan, United States*

Question: *is it tomato a fruit or a vegetable?*
Answer*(short: knowledge)*: *A tomato is a fruit.*

Question: *what to do if someone has a febrile seizure?*
Answer*(collection)*: *['Place her on the floor or bed away from any hard or sharp objects.', 'Turn her head to the side so that any saliva or vomit can drain from her mouth.', 'Do not put anything into her mouth; she will not swallow her tongue.', "Call your child's doctor."]*

Question: *how to check who saw your facebook story?*
Answer*(collection)*: *['Go to the Stories section at the top of your News Feed.', 'Click Your Story.', "Your story viewers will be listed below Story Details to the right. If you don't see this, no one has viewed your story yet."]*

Question: *how to get a red light ticket dismissed?*
Answer*(collection)*: *["Know the Law. You can't expect to prepare an adequate defense without some knowledge of the traffic code.", 'Know Your Driving Record. ... ', 'Request a Deferral. ... ', 'Tell a Convincing Story. ... ', 'Challenge the Traffic Cameras. ... ', 'Defensive Driving Course.']*

Question: *what to do when your toddler keeps crying?*
Answer*(collection)*: *['If you think your child might be tired, a rest might help. ... ', 'If the crying happens at bedtime, you might need some help settling your child.', 'If your child is angry or having a tantrum, take him somewhere safe to calm down.', 'If your child is frustrated, try to work out a solution together.']*

Question: *what are the disadvantages of using quantitative research methods?*
Answer*(collection)*: *['collect a much narrower and sometimes superficial dataset.', 'results are limited as they provide numerical descriptions rather than detailed narrative and generally provide less elaborate accounts of human perception.']*

Question: *what are holy places in christianity?*
Answer*(collection)*: *['Sephoria, where the Virgin Mary was said to have spent her childhood.', "The River Jordan, site of Christ's baptism.", 'Cave dwelling of John the Baptist.', 'Syria.', 'Galilee (North Israel/South Lebanon)', 'Sea of Galilee.']*

Figure 9: More examples from GooAQ. Instances of questions with the same type share background colors.