# 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 question-answering (QA) literature has failed to adequately address the answer diversity of questions. To this end, we present GooAQ, a large-scale dataset with a variety of answer type. This dataset contains with over 5 million questions and 3 million answers collected from Google. GooAQ *questions* are collected semi-automatically from the Google search engine using its autocomplete feature. This results in naturalistic questions of practical interest that are nonetheless short and expressed using simple language. GooAQ *answers* are mined from Google's responses to our collected questions, specifically from the answer boxes in the search results. This yields a rich space of answer types, containing both textual answers (short and long) as well as more structured ones such as collections. We benchmark T5 models on GooAQ and observe that: (a) in line with recent work, LM's strong performance on GooAQ's short-answer questions heavily benefit from annotated data; however, (b) their quality in generating coherent and accurate responses 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. 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, by using background knowledge either stored in the QA model or retrieved from large corpora or knowledge bases. Open QA datasets, however, focus mainly on short responses (Berant et al., 2013; Joshi et al., 2017; Lee et al., 2019; Lewis et al., 2021; Bhakthavatsalam et al., 2021), and so do the models designed for them (Roberts et al., 2020; Lewis et al., 2020b; Lee et al., 2020; Min et al., 2021; Lewis et al., 2021). Further, the short responses considered typically inquire about people, dates, and counts (e.g., 65% of Natural Questions (Kwiatkowski et al., 2019) begin with 'who', 'when', or 'how many'; cf. Fig 2)).

In contrast, many of the everyday questions that humans deal with and pose to search engines have a more diverse set of responses, 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', '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 richer types, e.g., unit conversion, time zone conversion, or various kinds of knowledge look-up ('how much', 'when is', etc.). Such answer type diversity is not represented in any existing dataset.

We propose GooAQ, the first open QA benchmark containing *questions with all of the above answer types within a unified dataset, collected using the same, coherent process.* Specifically, GooAQ contains 3 million questions with short, snippet, or collection answers, such as the ones shown in Fig. 1.

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 use as the gold answer.

Having a variety of answer types within a

[1] https://github.com/allenai/gooaq

**Question:** *how many calories burned 30 minutes crossfit?*
**Answer***(short: from the snippect)*: *260 calories*
**Answer***(snippet)*: *According to the American Council on Exercise, a 115-pound person running for 30 minutes at a slow-to-moderate pace (a 10-minute mile) would burn about* ***260 calories****.*

**Question:** *what is an assignment what is the difference between an assignment and a delegation?*
**Answer***(snippet)*: *The difference 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 is the gravitational force of uranus?*
**Answer***(short: knowledge)*: *8.87 m/s²*

**Question:** *what is the square feet of an acre?*
**Answer** *(short: unit-conversion)*: *43560 Square foot*

**Question:** *what is the time difference between south africa and mauritius?*
**Answer***(short: time-conversion)*: *Mauritius is 2 hours ahead of South Africa*

**Question:** *what are the ingredients used in making black soap?*
**Answer***(collection)*: *[9 oz Coconut Oil., 20 oz Palm Oil., 3.5 oz Shea Butter., 0.6 oz Coconut Carbon., 0.5 - 1.5 oz Fragrance or Essential Oil., 14 oz Water.]*

**Question:** *what are the steps for decision making?*
**Answer***(collection)*: *[Step 1: Identify the decision You realize that you need to make a decision. , Step 2: Gather relevant information. Step 3: Identify the alternatives. Step 4: Weigh the evidence. Step 5: Choose among alternatives. Step 6: Take action. Step 7: Review your decision & its consequences.]*

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.

single, coherent, open QA benchmark enables a quantitative study of the inherent differences across them. Specifically, we use GooAQ to ask whether models 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?*

We explore these questions in the context of generative pre-trained language models (LMs) (Lewis et al., 2020a; Raffel et al., 2020) as *self-contained reasoners*, without explicit access to additional retrieved information. In particular, we benchmark the powerful T5 language model (Raffel et al., 2020) on GooAQ, with both automatic metrics and human evaluation (§4).

To study ($Q_1$), we train models (separately for each response type) with little labeled data ($2k$ questions), mainly to convey the nature of the task. While LMs struggle, as expected, in this setting on *short* response questions, they perform surprisingly well in generating *snippet* and *collection* responses (e.g., humans prefer T5-11B's response to Google's response in 30% of such questions; Fig. 4, bottom-right plots). 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. On ($Q_2$), 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. On ($Q_3$), larger models, as expected, are more effective in all cases, but the gains are much more pronounced for *snippet* and *collection* response generation (20+%) as compared to *short* responses (5-10%), under human evaluation.

We hope GooAQ will facilitate research towards improving models to answer snippet and collection response questions. While the largest models we consider achieve surprisingly high scores on these questions, they still lag behind gold responses in both automated and human evaluations. Importantly, due to little benefit observed from more annotated data on such questions, human parity requires rethinking the approach towards better models.

We find GooAQ to be a valuable resource for training models. E.g., T5 trained on our snippet and collection questions shows strong zero-shot generalization to ELI5 (Fan et al., 2019), a long-answer dataset, achieving a score within X% of the state-of-the-art models trained using ELI5 data.

**Contributions.** (a) 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. (b) We benchmark state-of-art models on GooAQ, both in terms of automatic and human judgments, and observe remarkable differences in how models behave w.r.t. various response types. (c) We demonstrate that GooAQ is also a valuable model training resource by showing strong zero-shot generalization to ELI5 (Fan et al., 2019). We hope this datatset will spur further research into open QA under diverse response types.

## 2 Related Work

A closely related work is the Natural-Questions (NQ) dataset (Kwiatkowski et al., 2019; Lee et al., 2019) which contains common questions written by Google users. While our questions (extracted via autocomplete) likely approximate questions commonly asked on Google, we show that our dataset represents a wider distribution of questions (§3.2), 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. 2), this is not the case for GOOAQ.

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 much 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; Joachims et al., 2017). Although academic researchers do not have direct access to information collected from the users of search engines, the data gathered from 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 start by describing how GOOAQ was collected, followed by key dataset statistics and quality assessment.

## 3.1 Dataset Construction

### 3.1.1 Query Extraction

To extract a rich yet natural set of questions we use Google auto-completion.[2] As noted above, 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 are highly reflective of 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 rely on 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 contains the answer to a given question, and in some cases, the answer box shows the answer directly.

We first scrape the search results for all of our questions (§3.1.1). This is the main extraction bottleneck, which was done over a span of 2 months. Subsequently, 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.

After the answer extraction step, we have all the necessary information to create a question in GOOAQ, such as the examples in Fig. 1.[3]

## 3.2 Data Statistics

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

To better understand the content of GOOAQ, we present several distributions from the data. Fig. 3 shows the length distribution of the GOOAQ questions and that of Natural-Questions (Kwiatkowski

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

[3] We define 'short' response questions to be the union of 'knowledge', 'unit-conversion', 'time-conversion' and short answers from the 'snippet' responses (cf. Fig. 1).

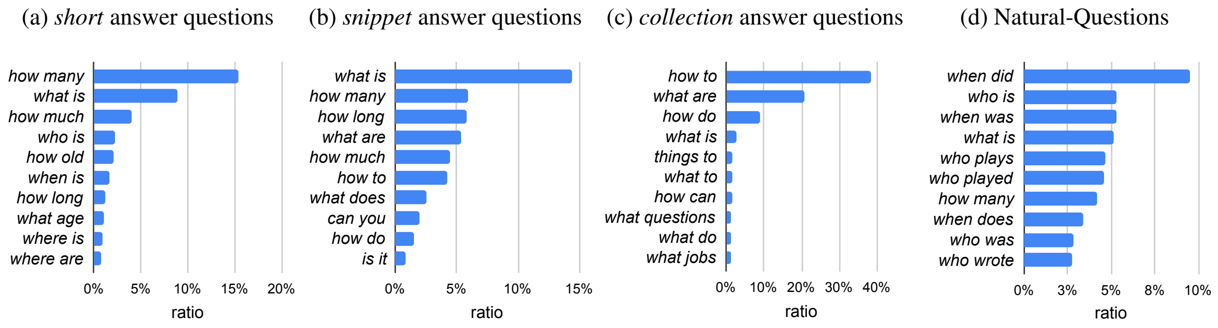

(a) *short* answer questions (b) *snippet* answer questions (c) *collection* answer questions (d) Natural-Questions

Figure 2: The distribution of common bigrams in questions of GooAQ (a,b,c) vs. Natural-Questions (d).

| Statistic | Value |
|---|---|
| # of questions | 5.0M |
| # of questions w/ answers | 3.1M |
| # of 'snippet' questions | 2.7M |
| ↳ the subset with 'short' answers | 196k |
| # of 'collection' questions | 329k |
| # of 'unit conversion' questions | 45k |
| # of 'knowledge' questions | 32k |
| # of 'time conversion' questions | 2.5k |

Table 1: GooAQ statistics

et al., 2019). While a vast majority of NQ questions contain 8-10 tokens, GooAQ questions have a somewhat wider length span.

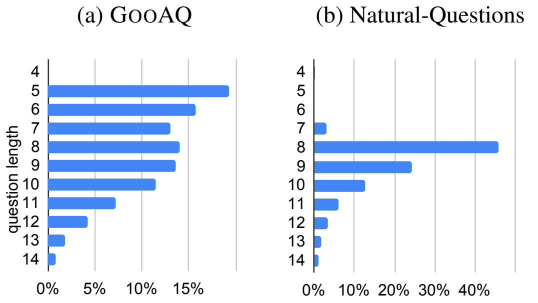

(a) GooAQ (b) Natural-Questions

Figure 3: Comparison of question length distributions

To better understand the type of questions, we show in Fig. 2 the distribution of the most frequent opening bigrams of the questions. 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 patterns '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. 2; 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.

### 3.3 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 clear and complete.

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 the annotators to avoid using Google search (which is what we used to we mined GooAQ) when annotating the quality of shown instances.

We compute aggregate measurements 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 2, 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.

| type | % of valid questions | % of valid answers |
|---|---|---|
| knowledge | 96.3 | 74.2 |
| time-conv | 93.3 | 66.4 |
| unit-conv | 96.9 | 83.1 |
| snippet | 98.4 | 67.6 |
| snippet (short) | 98.7 | 84.5 |
| collection | 99.7 | 91.7 |
| avg | 98.3 | 77.9 |

Table 2: Summary of GOOAQ quality evaluation by crowdworkers. 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.

The indicates a reasonable quality of GOOAQ question-answer pairs, as evaluated directly in isolation from any systems.

## 4 Experimental Setup

GOOAQ naturally forms a dataset for the task of open QA, where the input is a question and the output is the answer to that question. Unlike the reading comprehension setting, the context for answering the question is not provided as part of the input. Further, we consider the so-called 'closed-book' setup (Roberts et al., 2020) 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.

**Setup.** We split GOOAQ into three sub-tasks: ($\mathcal{T}_{short}$) *short* responses questions (cf. footnote 3). ($\mathcal{T}_{snippet}$) *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 from each question type. We experiment with variable training data size to better understand the value of the labeled data. Prior work has shown that leakage from training data to the evaluation sets often result in unrealistically high scores (Lewis et al., 2020b). 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.

**Models.** For our evaluation, we use the T5 model (Raffel et al., 2020), a recent text-to-text framework that has achieved state-of-art results on a variety of tasks, including open QA (Roberts et al., 2020). 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.

**Automatic evaluation metrics.** We use ROUGE-L (Lin, 2004) as it is a common metric for text generation tasks. The results of the automatic evaluation for each sub-task are shown in the top row of Fig. 4. For short answer questions, we show the automatic evaluation of each sub-type (unit conversion, time conversion, etc.) in Fig. 6.

**Human evaluation.** Automatic evaluation is well-known to be sub-optimal for text generation models, especially for larger text. Hence, in addition to the automatic metric, we also perform human evaluation of the generated responses, using the judgments of AMT crowdworkers. Specifically, we ask the crowdworkers to indicate their preferred answers (among gold answer and the model prediction). The annotation interface is shown in Fig. 5, which is essentially the same template used for the quality assessment of the dataset (§3.3), except that here the crowdworkers are shown a *pair* of responses for each question—the gold answer and the one generated by the model—turning the task into a *comparative* one.

Before annotating every instance, we remind the annotators to avoid using Google. Then we ask them to check if the provided question is clear enough and it makes sense. Upon indicating 'yes' to question quality, they are shown two answers labeled 'A' and 'B' (one Google's answer and one generated by our models, ordered randomly). We

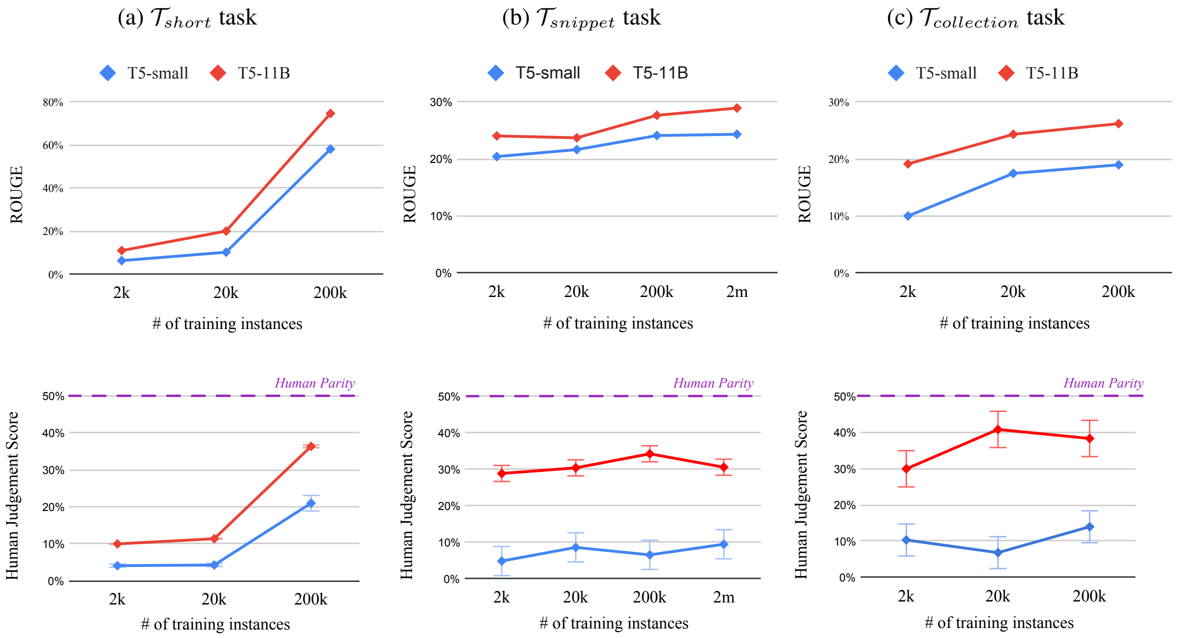


Figure 4: Evaluation of T5 (small,11B) models on different sub-tasks of GooAQ via *automatic* metrics (top) and *human* judgements (bottom). For human evaluation, 50% is the border at which the model output and the ground truth responses are indistinguishable. The short-answer sub-tasks ($\mathcal{T}_{short}$; left) have a relatively low performance when supervised with $2k$ instances. However, benefit more then the long-answer sub-tasks ($\mathcal{T}_{snippet}$ & $\mathcal{T}_{collection}$) from availability of more labeled data. On contrary, long-answer sub-tasks benefit very little 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).

ask them to indicate the answer they prefer (if any).

Each question is annotated by 5 independent annotators and aggregated via a majority vote of the 5 labels. If annotators consistently prefer model predictions, we assign 1 credit to the prediction. Otherwise, the model receives 0 credit. To compute an overall accuracy score for a given model, we average the instance-level scores, after discarding the questions indicated as invalid ('this question makes no sense').

The resulting *human-evaluation* metric indicates the percentage of the cases where model predictions were preferred over ground-truth answers. In this evaluation, 50% is the margin where the annotators are not able to distinguish the model's responses from the ground-truth responses (Google's answers) in any meaningful way. The results of human evaluation are shown in the bottom row of Fig. 4.[4]

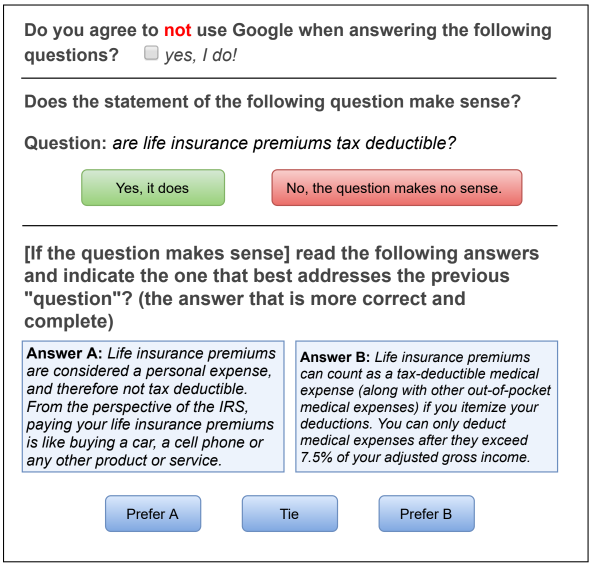
## 5 Empirical Results

### 5.1 Main Results

**On long-answer sub-tasks, pre-training of the models carries most of the weight.** Both automatic and human evaluations (Fig. 4; middle & right), indicate that the model performances are quite robust to the increase in the number of labeled data. Essentially, the ability of language-model in addressing such questions is mainly enabled by their pre-training, as indicated by their performance when they're supervised with only $2k$ instances.

To understand this observation, one needs to put into perspective several factors that are at play. First, unlike short answer questions which typically ask for encyclopedic knowledge and therefore, *correctness* of the answers matter the most, we suspect that in snippet and collection questions, *coherence* of the responses carry heavier weight. This is partly due to the nature of the collected questions which typically refer to high-level notions that can be answered in a variety of ways. For example, the snippet response to the question of *how many calories burned 30 minutes crossfit?*

[4] The templates used for human evaluation are available (cf. Footnote 1).



**Do you agree to not use Google when answering the following questions?** □ *yes, I do!*

**Does the statement of the following question make sense?**

**Question:** *are life insurance premiums tax deductible?*

**[If the question makes sense] read the following answers and indicate the one that best addresses the previous "question"? (the answer that is more correct and complete)**

**Answer A:** *Life insurance premiums are considered a personal expense, and therefore not tax deductible. From the perspective of the IRS, paying your life insurance premiums is like buying a car, a cell phone or any other product or service.*

**Answer B:** *Life insurance premiums can count as a tax-deductible medical expense (along with other out-of-pocket medical expenses) if you itemize your deductions. You can only deduct medical expenses after they exceed 7.5% of your adjusted gross income.*

Figure 5: Crowdsourcing interface used for our human evaluation.

(Fig. 1) could refer to a range of calorie consumption, depending on the choice of activity during crosssfit or the attributes of the person working out (and all of these responses would be equally correct.)

**Labeled data is more helpful to the short answer questions.** Based on the automatic evaluation (Fig. 4; top-left) the performance of both models on the short response question quickly improves as we increase the number of the training data (the opposite of long-response questions).

The breakdown of the automatic evaluation for different types of short response questions is shown in the top row of Fig. 6. As expected, certain niche question types (such as 'unit-conversion') benefit the most from labeled data. On the other hand, open-ended question types (such as 'knowledge' lookup) benefit less from more labeled data.

**Human evaluation accentuates the gap between '11B' and 'small' models, especially on long-response questions.** This is compatible with the recent work (Min et al., 2021) where we also see that the gap between two reasonably different systems is bigger in terms of human evaluation (Fig. 4, bottom row compared top row). We hypothesize this is due to the crudeness of automatic evaluation metrics, and an indication of human evaluation's necessity to distinguish between the nuanced differences between the generated responses. What is more interesting (and not noticed in the prior work) is that, the gap between automatic-human evaluation is bigger for our snippet and collection questions that involve longer responses. This is, at least partly, due to the inaccuracy of automatic metrics in evaluating long text generation.

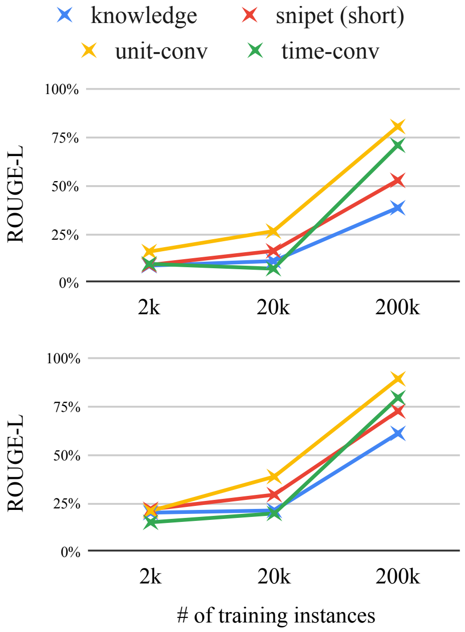


Figure 6: *Automatic* evaluation of T5 (small: top, 11B: bottom) models on different types of the questions included in short-answer sub-task ($\mathcal{T}_{short}$). 'unit-conversion' questions benefit the most from more labeled data, while 'knowledge' lookup questions are the opposite.

**Few-shot '11B' models achieve high performance, but not yet comparable with the gold annotations.** As mentioned earlier, our human evaluation measures the comparative quality of the model predictions and the ground truth responses (Google's answers). Hence, a value of 50% in this evaluation is an indication that the predictions are on par with (indistinguishable from) the ground-truth responses.

As can be seen in the bottom row of Fig. 4, while even this large model is still not on par with Google's gold responses, it comes quite close to it even when the model has seen a small amount of labeled instances. We hope this gap will encourage further research in building stronger models.

## 5.2 Error Analysis

To gain better intuitions about the mistakes made by the models, we conducted a small-scale errors analysis of the model predictions. For each model, we (one of the authors) annotated 30 predictions predictions, and labeled them with the following error categories that are inspired from the existing evaluations of 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 in the predictions.

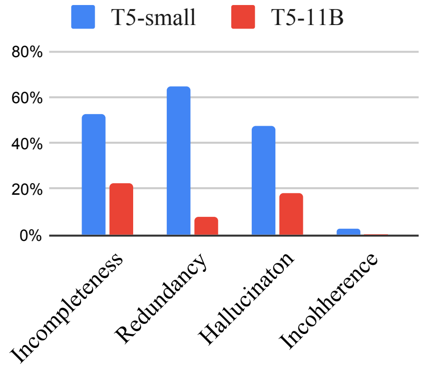


Figure 7: The distribution of the errors across different models

The results of our error analysis are shown below in Fig. 7. As expected the 'small' models have higher error percentages across all categories. They particularly suffer from a lot of *redundancy* and *incompleteness*. Overall, all the models have very little *incoherence* mainly because this category is directly addressed during the model pre-training.

### 5.3 Extrinsic Utility of GooAQ

We assess the value of GooAQ on a slightly different distribution of questions. In particular, we evaluate our models on ELI5 (Fan et al., 2019), a recent dataset for long-answer question-answering extracted from Reddit posts.

Our evaluation on ELI5 (summarized in Table 3) show that, our models our T5 models trained on GooAQ (snippet-answer subset) perform quite well when evaluated on ELI5, *better than* the same architectures trained with ELI5 training set and on par with the state-of-the-art models that use retrieval engines.

We hypothesize that, despite the fact that GooAQ is collected differently than ELI5, there is a good portion of ELI5 questions that are covered by GooAQ, indicating its good coverage of common questions posed by ordinary users.

| Model | Uses retrieval? | Score |
|---|---|---|
| T5-small (GooAQ; snippet; *2M*) | *no* | 21.8 |
| T5-11B (GooAQ; snippet; *2M*) | *no* | **22.9** |
| T5-small (ELI5) | *no* | 19 |
| T5-11B (ELI5) | *no* | 22.7 |
| RAG * (Krishna et al, 2021) | *yes* | 14.1 |
| RT + REALM * (Krishna et al, 2021) | *yes* | 23.4 |

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

## 6 Discussion

**Knowledge leakage in the evaluation.** One recent finding in the field is about knowledge leakage between train and evaluation sets (Lewis et al., 2020b; Emami et al., 2020). Similar concerns has motivated our careful train/evaluation splits (§4) and experiments with varying training sizes. Nevertheless, we found it challenging to define (and assess) the amount of leakage from the training data to evaluation. We delegate and welcome such studies on our dataset.

**Are we mimicking Google's QA?** A reader might question the value of the study by citing that the website from which these annotations were crawled, had likely also used a QA system for such extraction. So, this work might raise eyebrows since our work might seem like we are reverse-engineering Google's internal QA system. While we (the authors) are not aware of how Google answer boxes work, we suspect that it is much more than a QA system with the current AI technology. Such a system likely makes heavy use of the implicit user feedback (billions of click information, the structure of the web links, etc.), in addition to all the explicit feedback from Google users. In sum, the collected data from answer boxes captures various signals that have participated in curating high-quality answers.

**Replicating human evaluation.** One challenge with respect to the progress on long-form QA task is the evaluation of responses. To facilitate the future work on GooAQ and replicability of our human evaluations, we have released the templates we have used to crowdsource human judgements. There is also actively proposals streamlining evaluation of text generation tasks (Khashabi et al.,

2021) which we might adopt, if there is enough interest in the task.

**Scope of our conclusions.** It is worth emphasizing that one must be careful in taking these conclusions out of the context of this study (i.e., the dataset at hand, the models, evaluation metrics, etc.). While we hypothesize that these findings are relatively general, one might be able to come up with a different long-form QA dataset on which the same baselines show a wildly different behavior.

## 7 Conclusion

In this work, we study QA under diverse types. For this purpose, we collected GOOAQ, an enormous collection of question-answer pairs collected from Google with a variety of (short and long) types, all of which are collected with a unified process. The questions are collected from the autocomplete system which likely reflects a *natural* distribution of questions asked on the search engine. We benchmark a state-of-the-art self-contained text generation models (with no retrieval) on three different sub-tasks of GOOAQ (short, snippet and collection), and evaluate via both automatic and human evaluation. Our analysis indicates the distinct behavior of LMs on long and short response questions: while short response models benefit heavily from more labeled data, long responses are mostly driven by pre-training of the models. We hope these analyses and the released data will benefit our understanding of the limits and capabilities of QA systems in dealing with various forms of responses.

## Acknowledgement

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

## 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*.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine.

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*, volume 51, 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*.

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.

Nayeon Lee, Belinda Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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. 2020b. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *arXiv preprint arXiv:2102.07033*.

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 used for query

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 Error Analysis

To gain better intuitions about the mistakes made by the models, we conducted a small-scale errors analysis of the model predictions. For each model, we (one of the authors) annotated 30 predictions, and labeled them with the following error categories that are inspired by the existing evaluations of summarization (Chaganty et al., 2018):

- Incompleteness: the prediction has some relevant points but not as detailed as expected.
- Redundancy: the prediction ha repeated content.
- Hallucination: the prediction is making up new facts.
- Incoherence: the prediction lack of fluency and coherency.

Here are examples of such errors in model predictions:

**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.)

The results of our error analysis are shown below in Fig. 7. As expected the ‘small’ models have higher bars across all error categories. They particularly suffer from a lot of ‘redundancy’ and ‘incompleteness’. Overall, all the models have very little ‘incoherence’ mainly because this category is directly addressed during the model pre-training.

**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 8: More examples from GOOAQ. Instances of questions with the same type share background colors.