Large language models are increasingly capable of generating fluent-appearing text with relatively little task-specific supervision. But can these models accurately explain classification decisions? We consider the task of generating free-text explanations using a small number of human-written examples (i.e., in a few-shot manner). We find that (1) higher-quality, human-authored prompts result in higher quality generations; and (2) surprisingly, in a head-to-head comparison, humans often prefer explanations generated by GPT-3 to crowdsourced explanations in existing datasets. Our human studies also show, however, that while models often produce factual, grammatical, and sufficient explanations, they have room to improve along axes such as providing novel information and supporting the label. We create a pipeline that combines GPT-3 with a supervised filter that incorporates binary acceptability judgments from humans in the loop. Despite significant subjectivity intrinsic to judging acceptability, our approach is able to consistently filter GPT-3 generated explanations deemed acceptable by humans.

1 Introduction

As natural language understanding tasks have become increasingly complex, the field of explainable natural language processing (exNLP) has embraced explanations written in free-form natural language. In contrast to extracting input highlighting explanations, free-text rationales provide a natural interface between machine computation and human end-users (Hendricks et al., 2016; Camburu et al., 2018). The dominant paradigm for producing free-text explanations is via direct supervision, i.e., training an autoregressive, generative language model to predict human-authored explanations directly (Kim et al., 2018; Park et al., 2018; Ehsan et al., 2018; Narang et al., 2020; Wiegreffe et al., 2021, i.a.).

However, collecting high-quality, human-written explanations to serve as supervision is difficult and expensive. More than 70% of existing free-text explanation datasets are crowdsourced (Wiegreffe and Marasović, 2021), and even the most meticulous crowdworking efforts frequently fail to elicit logically consistent and grammatical explanations (leading downstream to poor-quality models) (Narang et al., 2020). Furthermore, a lack of standardized crowdsourcing design has resulted in highly varied datasets, which are hard to compare or combine (Tan, 2021).

Recent progress in few-shot learning provides
a potentially promising alternate to large-scale crowdsourcing. The in-context learning paradigm, wherein powerful language models are prompted in a few-shot manner with just a few examples, has proven surprisingly effective across a range of NLP tasks (Radford et al., 2019; Brown et al., 2020; Shin et al., 2020; Schick and Schütze, 2021a, i.a.). In this work, we ask: can language models also generate reliable explanations? We present a human subjects study with a surprising finding: in-context learning with GPT-3 (Brown et al., 2020) produces explanations competitive with crowdsourced explanations in existing datasets (§2).

Two additional human subjects studies, however, demonstrate that GPT-3-generated explanations still have significant room for improvement along axes such as providing new information (i.e. avoiding repetition) and supporting the label. In particular, human subjects found less than half of greedy-decoded GPT-3 generated explanations to be acceptable with high agreement.

To further improve generation quality, we reframe the role of crowd annotators: instead of authoring explanations as in prior work, we (1) repeatedly query GPT-3 to generate multiple candidate explanations for each input instance; and (2) ask crowdworkers to rate the acceptability of each candidate generation. After showing that GPT-3 can usually generate an explanation that humans find acceptable within as few as five queries (§3), we use a small number of these binary crowdworker judgments to supervise an acceptability filtering model, which can be applied to select high quality candidates among GPT-3’s outputs (Figure 1; §4).

Despite intrinsic subjectivity in acceptability ratings, our supervised model improves upon the already-competitive few-shot paradigm by consistently selecting (human-identified) high quality explanations better than strong baselines. Human evaluations reveal that the filtration model not only improves acceptability, but also, other axes like supporting the label and providing novel information.

In summary, our main findings are:

i. in-context learning with GPT-3 produces surprisingly competitive explanations, providing a promising alternative to crowd-authored free-text explanation corpora;

ii. binary human labeling can instead be leveraged to train a filter that selects high-quality machine-generated explanations; and

iii. in areas where GPT-3 struggles, including information content, supporting the label, and overall acceptability, our proposed overgenerate-and-filter pipeline improves generated explanations.

We publicly release our code and human-annotated data.¹

2 In-Context Learning is Competitive with Crowdsourced Datasets

Is in-context learning with GPT-3 a viable strategy to generate free-text explanations? To this end, we investigate three research questions:

- Are GPT-3-generated explanations preferable to crowdsourced ones in existing datasets? (§2.1)
- Can improving prompt quality improve GPT-3-generated explanations? (§2.2)
- Along what fine-grained dimensions are GPT-3-generated explanations preferred? (§2.3)

Explanation tasks and datasets. We consider two English tasks: CommonsenseQA and natural language inference (NLI), shown in Table 1. CommonsenseQA (Talmor et al., 2019) is a multiple choice task posed over commonsense questions. Crowdsourced free-text explanations for instances in CommonsenseQA are provided in the CoS-E v1.11 (Rajani et al., 2019) and ECQA (Aggarwal et al., 2021) datasets. ECQA explanations are counterfactual, i.e., annotators were instructed to explain not only the correct answer choice but also why the others are incorrect.² ECQA was released to address the quality issues of CoS-E (Narang et al., 2020); we consider both to provide perspective on the impact of prompt quality. Our second task is NLI, which involves inferring whether a given hypothesis sentence entails, contradicts, or is neutral towards a premise. This task is instantiated with the SNLI dataset (Bowman et al., 2015), and crowdsourced free-text explanations are collected in the e-SNLI dataset (Camburu et al., 2018). For each task, we report results on a fixed, randomly-sampled 250-instance test set not observed during prompt design.

In-context learning for explanations. We use GPT-3 Davinci (Brown et al., 2020), an autoregressive language model with ~175B parameters trained on a large dataset of text scraped from the

¹https://anonymous/
²We do not perform counterfactual human evaluations; ECQA explanations are thus often viewed by annotators as having “too much” information, see §2.3.
We use a total of 115 randomly sampled train instances with greedy decoding. More details about prompt construction are in Appendix A; example prompts are given in Tables 8-9.

For simplicity, because we aim to focus on free-text explanations, we assume access to the gold label. In early experiments where we also had GPT-3 explanations generated under different conditions (“head-to-head”), we then ask them to make a preferential selection on a 5-point Likert scale, collecting 3 annotations per data point. Appendix B contains further details. We report inter-annotator agreement using Krippendorff’s $\alpha$ (Krippendorff, 2011).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Preferred Explanation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold-Standard</td>
<td>Tie</td>
</tr>
<tr>
<td>CoS-E</td>
<td>20.3</td>
</tr>
<tr>
<td>ECQA</td>
<td>52.7</td>
</tr>
<tr>
<td>e-SNLI</td>
<td>63.6</td>
</tr>
</tbody>
</table>

Table 2: Head-to-head human evaluation (3 annotations each) for 250 explanations generated by GPT-3 vs. written by crowdworkers in the datasets along with Krippendorff’s $\alpha$. Results are shown as % preferences. GPT-3 explanations were prompted with explanations from the respective datasets.

Crowdsourcing evaluation. Given that existing automatic metrics do not correlate well with human judgements of explanation quality (Clinciu et al., 2021; Kayser et al., 2021), we perform human-subjects studies on the Amazon Mechanical Turk (AMT) platform. We compensate crowdworkers at a rate of $15/hour. We present participants with a dataset instance, gold label, and two explanations for the instance generated under different conditions (“head-to-head”). We then ask them to make a preferential selection on a 5-point Likert scale, collecting 3 annotations per data point. Appendix B contains further details. We report inter-annotator agreement using Krippendorff’s $\alpha$ (Krippendorff, 2011).

2.1 RQ1: Are GPT-3 explanations preferred over crowdsourced ones?

We perform a head-to-head comparison of explanations generated by GPT-3 with greedy decoding vs. gold human-written explanations in the original datasets. The crowdsourced explanations serve as a reasonable upper bound for what a supervised explanation generation model trained on them could produce. Results are shown in Table 2. Some examples of GPT-3-preferred explanations are given in Table 1.

For CoS-E, the GPT-3 greedily-decoded explanations are frequently preferred or comparable to crowdsourced explanations, which is not too surprising for CoS-E, which has many ungrammatical explanations (Narang et al., 2020). And, while ECQA and e-SNLI explanations are strongly preferred to GPT-3, there are still a non-trivial number of cases where GPT-3 explanations are competitive (36.4% and 24.8%, respectively).

2.2 RQ2: Can improving prompt quality improve GPT-3-generated explanations?

Given that low-quality training instances may result in low-quality predictions (especially in a few shot setting), we ask: can we improve GPT-3 generations simply by conditioning on higher-quality instances? For prompt construction, we replace the 115 crowdsourced explanations from the original datasets with explanations carefully written by the

Table 1: Task-specific instances, along with their crowdsourced explanations from the respective datasets, shown alongside explanations generated greedily by GPT-3. In our experiments, the SNLI GPT-3 explanation was preferred over its corresponding e-SNLI explanation by 2/3 annotators. For CommonsenseQA, 3/3 preferred the GPT-3 explanation to the CoS-E explanation, and 2/3 to the ECQA one.

| Premise: Dark-haired man wearing a watch and oven mitt about to cook some meat in the kitchen. |
| Hypothesis: A man is cooking something to eat. |
| Label: entailment |
| e-SNLI (Camburu et al., 2018): Meat is cooked in a kitchen, and is a food that you eat. Using an oven mitt implies you’re about to cook with hot utensils. |
| GPT-3: Cooking is usually done to prepare food to eat. |
| CommonsenseQA (Talmor et al., 2019): |
| Question: What is the result of applying for job? |
| Answer Choices: anxiety and fear, increased workload, praise, less sleep, or being employed |
| Correct Choice: being employed |
| CoS-E (Rajani et al., 2019): being employed applying for job |
| ECQA (Aggarwal et al., 2021): Applying for a job is followed by attending interview which results in being employed. Applying for a job may not result in the other options. |
| GPT-3: Applying for a job can result in being employed, which is a positive outcome. |

For CoS-E, the GPT-3 greedily-decoded explanations are frequently preferred or comparable to crowdsourced explanations, which is not too surprising for CoS-E, which has many ungrammatical explanations (Narang et al., 2020). And, while ECQA and e-SNLI explanations are strongly preferred to GPT-3, there are still a non-trivial number of cases where GPT-3 explanations are competitive (36.4% and 24.8%, respectively).
authors of this paper (see Table 10 for examples). Our prompts are used to generate a different set of GPT-3 explanations on the same test data.

We perform a head-to-head human evaluation of the GPT-3 generations conditioned on the explanations we authored vs. those conditioned on the gold crowdsourced explanations generated in §2.1. Results in Table 3 show that, for all three corpora, generations conditioned on our explanations outperform generations conditioned on crowdsourced explanations. This effect is especially pronounced for CommonsenseQA; GPT-3 even reproduces known data artifacts in the CoS-E corpus when prompted with explanations from it (such as the phrase “rivers flow through valleys”, which appears 10 times in the prompt set). We repeat the experiment of §2.1, but with our prompts instead of dataset prompts.

With this change, GPT-3 generations are even more competitive (Table 4). For all three datasets, more than half the time, in-context learning results in an explanation at least as good as a human written explanation. These results provide evidence that in-context can enable end-user control over explanation format and quality by authoring as few as 8-24 examples. For subsequent experiments, we use the explanations written by the authors as prompts.

2.3 RQ3: What types of explanations does GPT-3 output?

Pairwise evaluations can only offer perspective on the relative quality of generated explanations. Are crowd annotators simply comparing explanations on surface-level features like grammaticality?
To understand finer-grained characteristics of explanations, we design a second human study to collect absolute Likert-scale judgments across seven axes of quality (with each explanation judged by 3 annotators). The first three axes capture surface-level features: generality, grammaticality, and factuality. The next three capture richer aspects of explanation quality: whether new information is introduced (a requirement for non-vacuous explanations), whether explanations support the gold label, and whether the amount of information given is sufficient. Finally, we ask for an overall judgement of quality: is the explanation acceptable? We explain our design process in Appendix B.2. Results on the crowdsourced and GPT-3 explanations for both tasks are given in Figure 2.4

For both tasks, GPT-3 explanations do well in all 3 surface-level categories, with statistically significantly greater ratings in generality and grammaticality (and factuality for CommonsenseQA) compared to crowdsourced explanations, and distributional means close to 1. In these categories, there is little room for improvement.

On the other hand, GPT-3 explanations do not contain as much new information as ECQA and e-SNLI explanations, and have substantial room to improve (mean=0.1 for both tasks compared to 0.6 for ECQA and 0.2 for SNLI; these differences are statistically significant at \( p \leq 0.01 \)). GPT-3 explanations are substantially more supportive of the label over CoS-E, but not as supportive as ECQA or e-SNLI (all statistically significant at \( p \leq 0.1 \)). Indeed, the mean rating of GPT-3 explanations for label support is 0.5 for CommonsenseQA and -0.1 for NLI, demonstrating room for improvement. These axes are crucial to ensuring explanations are not vacuous and are on-topic. Finally, GPT-3 explanations are judged as acceptable at higher rates than CoS-E or ECQA explanations, but not e-SNLI explanations. Mean scores of 0.5 (CommonsenseQA) and 0.0 (NLI) indicate that GPT-3 explanations have room to improve overall.

3 Beyond Greedy Explanations

While GPT-3 explanations demonstrate strength across surface-level features, and are surprisingly competitive in head-to-head settings, they can still be improved. Borrowing our acceptability crite-

4 Means, standard errors, and Wilcoxon signed rank test results are in Table 16; Krippendorff’s α is 0.48 for CommonsenseQA annotations and 0.31 for SNLI — see Table 14.

3.46

4 Improving Explanation Generation with Acceptability Filtering

The challenge of overgeneration is that GPT-3 alone cannot discern which of its samples are acceptable. We explore training a supervised filter on the collected labels. Our key intuition is that by re-framing the role of annotators from explanation authors to binary judges, we can alleviate the need to collect a large-scale explanations dataset—the result is a simpler, cheaper, and easier crowdsourcing setup to administer (§4.1). Moreover, we find that the filter can be trained with a relatively small amount of binary human judgments (§4.2). Figure 1 presents an overview of our pipeline.

4.1 Acceptability Annotations

We generate train/validation sets by repeating the procedure of generating 1 greedy and 4 sampled explanations for 991 and 1K instances, respectively, of the CommonsenseQA and SNLI training sets. Combining these with the annotated test sets from previous experiments results in a dataset of 1241/1250 instances in a 72/8/20% train/val/test split.

5 In §4, we show that these are not upper-bounds caused by intrinsic subjectivity, and that they can be improved upon.

6 Appendix C provides statistics for the case where 2/3 is the acceptability threshold, with very similar findings.
ratio for each task. We again collect 3 binary acceptability ratings for each instance, resulting in ~6200 instance-explanation pairs and ~19k individual annotations per task. Table 11 contains statistics. In addition, to ensure that models trained on these corpora do not overfit, random annotations from a group of annotators who did not participate in any of our previous annotation tasks (“Test2”). Krippendorff’s α for all acceptability annotations is 0.34 for CommonsenseQA and 0.39 for SNLI (see Table 13).

While we evaluate at test-time with the schema that instances that 3/3 annotators deem acceptable are acceptable and all others (0/3, 1/3, 2/3) are deemed unacceptable, preliminary experiments show that treating 2/3 and 3/3 agreement instances as acceptable and 0/3, 1/3 instances unacceptable during training performs best on the 3/3 evaluation criterion at test-time. We also train a variant where we randomly select one annotation from the three as the gold label (“without human agreement”).

### 4.2 Acceptability Filter

Concretely, given the problem instance (e.g., premise/hypothesis for NLI), the gold label, and the (generated) explanation, the acceptability filter predicts whether the explanation is acceptable. We fine-tune two sequence-to-sequence architectures, T5-Large (Raffel et al., 2020) and T0-3B (Sanh et al., 2021). Each model is trained 5x with different random seeds. Further training details are given in Appendix D.

### Baselines

We train an explanation-only baseline, which receives as input only the explanation; similar baselines have been proposed for NLI (Poliak et al., 2018; Gururangan et al., 2018). These models represent the hypothesis that annotator ratings can be reconstructed with only surface features of the explanation candidates, e.g., grammaticality. We also consider a negative log-likelihood (NLL) baseline, which uses GPT-3’s estimated probability as the acceptability classification score.

This is a slightly more competitive baseline than greedy; greedy usually (but not always) produces the highest-likelihood explanation.\(^6\)

\(^6\)Our results don’t significantly change if a 2/3 cutoff is used at test time instead; Appendix E contains the results.

<table>
<thead>
<tr>
<th>“Select-1” Acc@3/3</th>
<th>Expl.-level AP@3/3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Random</td>
<td>26.5±2</td>
</tr>
<tr>
<td>Constant</td>
<td>—</td>
</tr>
<tr>
<td>NLL</td>
<td>41.8</td>
</tr>
<tr>
<td>T5-L Expl.-only</td>
<td>40.2±0.3</td>
</tr>
<tr>
<td>T0-3B Expl.-only</td>
<td>42.6±1.4</td>
</tr>
<tr>
<td>T5-L w/o HA</td>
<td>46.8±2.3</td>
</tr>
<tr>
<td>T5-L</td>
<td>46.4±2.9</td>
</tr>
<tr>
<td>T0-3B w/o HA</td>
<td>48.4±2.0</td>
</tr>
<tr>
<td>T0-3B</td>
<td>48.8±0.9</td>
</tr>
<tr>
<td>Oracle U.B.</td>
<td>78.0</td>
</tr>
</tbody>
</table>

Table 5: Results for acceptability classifiers trained on CommonsenseQA. Subscripts indicate standard error over models trained with 5 different random seeds. “w/o HA” = without human agreement. “Oracle U.B.” indicates upper bound based on dataset properties ($\S3$).

### 4.3 Evaluation

We consider three evaluation settings. The first is instance-level (“select-1”), where the system returns 1 explanation selected from the set of 5 for each instance. We return the explanation with the highest model-estimated probability and report instance-level accuracy, i.e., the % of instances for which a gold acceptable explanation is selected.

We also evaluate at the explanation-level, where we treat each explanation independently and compute metrics over the full dataset. This aligns with the binary classification training of the models (cross-entropy on the explanation labels) and is suited for the setting in which we want to return all of the acceptable explanations per instance. In this setting, we report average precision (AP), an estimate of area under the precision-recall curve.

Finally, we perform an absolute human evaluation ($\S2.3$) on the subset of instances where the filter model does not select the greedy explanation as the best, i.e., comparing “select-1” performance to a greedy baseline on the instances where it differs. For CommonsenseQA/SNLI, $n = 156/91$.

### 4.4 Results

Classifier performance is given in Tables 5-6.

**Effect of model size.** On CommonsenseQA, T0-3B outperforms T5-Large by ~2-4% select-1 accuracy and ~5-6% explanation-level AP across splits. We use T0-3B in subsequent experiments.

**NLL baseline vs. full model.** For both tasks on both validation and test sets, T0-3B outperforms the NLL baseline substantially. On CommonsenseQA, we observe a 7-8% gain in instance-
level accuracy, and a gain of 18% explanation-level AP on the test set. This provides strong evidence that the supervised model is able to incorporate binary human feedback to predict acceptable explanations at a level much higher that GPT-3 achieves on its own. Our filter model predicts a different explanation than NLL in the “select-1” setting for 42 out of 250 for NLI; we present examples in Table 7 and Table 12.

| Explanation only vs. full model. | Our results suggest that our models are leveraging feature interactions between the instance and explanation to make their predictions. Without instance-level context, the explanation-only baselines are on average more than 5 points worse across metrics. Though they underperform significantly relative to the full model, explanation-only baselines do fare surprisingly well, indicating that shallow features like factuality and grammaticality may represent latent
| factors in human acceptability judgments.

The effect of multiple training annotations. In some cases, performance improves if the training instances are labeled with the consensus of three annotators (vs. the singularly annotated case “w/o HA”), though the effects are not consistent. In most cases, using consensus agreement results in reduced variance across random seeds. However, given that training on consensus requires 3x the annotations, the gains may not outweigh the data collection effort.

Our model doesn’t overfit to specific annotators. Reassuringly, the performance of our model when evaluated on the NLI test set labeled by separate annotators (“Test2”) is comparable to the original test set (instance-level accuracy drops a few points, but explanation-level AP slightly rises). Importantly, we also reach the same conclusions on this test set regarding the superior performance of our model with respect to the baselines.

Our model improves generated explanations along desirable traits. We present our absolute human evaluation for greedy vs. filtered explanations from GPT-3 in Figure 3— for both tasks, explanations filtered by our model more readily introduce new information, support the label, and contain at least enough information for both tasks (in addition to being more acceptable). Interestingly, greedy explanations still prevail in surface-level features (grammaticality and, in the case of CommonsenseQA, factuality). All of these differences are statistically significant at small values of p (see Table 17). The differences in generality (and particular to explain the differences in human acceptability judgments.

Table 6: Results for SNLI explanation acceptability; all model results are on T0-3B. See Table 5’s caption.

![Figure 3: Absolute evaluation results in the “select-1” setting for the instances where our best-performing filter model does not select the greedy explanation (156 instances for CommonsenseQA (top); 91 for NLI (bottom)). See caption of Figure 2 and the Appendix-Table 17 for more details.](image-url)

<table>
<thead>
<tr>
<th>“Select-1” Acc@0/3</th>
<th>Explanation-level AP@0/3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
</tr>
<tr>
<td>Random</td>
<td>31.5 ± 0.9</td>
</tr>
<tr>
<td>Constant</td>
<td>31.2 ± 0.9</td>
</tr>
<tr>
<td>NLL</td>
<td>33.0 ± 0.9</td>
</tr>
<tr>
<td>Expl.-only w/o HA</td>
<td>38.3 ± 0.9</td>
</tr>
<tr>
<td>Full</td>
<td>37.8 ± 0.9</td>
</tr>
<tr>
<td>Oracle U.B.</td>
<td>51.0 ± 2.4</td>
</tr>
</tbody>
</table>
Table 7: Randomly-selected instances that our filter model predicts differently than NLL at the “select-1” task and got correct, but NLL got incorrect.

In summary. We have demonstrated the effectiveness of modeling binary crowd judgements of acceptability as a means to select candidates from GPT-3 which are deemed acceptable at a high agreement. For the method that does not leverage human agreement, this is done with only ~5k binary annotations. We additionally demonstrate that our filtered explanations improve upon greedy generations in fine-grained categories that probe their topical relevance and meaningful content. The gap between our best model and the upper-bound oracle indicates that there is still substantial room for improvement in both task settings. Future work may investigate sampling more explanations, or incorporating other sources of supervision signal.

5 Related Work

Free-text explanation generation. The earliest neural free-text explanation models did so for computer vision applications (Hendricks et al., 2016; Park et al., 2018; Kim et al., 2018) and NLI (Camburu et al., 2018). These methods relied on supervised datasets to train the explanation generator. Others have proposed to generate explanations or clarifications to improve task performance in a supervised (Rajani et al., 2019) or unsupervised (Shwartz et al., 2020) manner. Yordanov et al. (2021) study transfer learning between datasets for few-shot generation.

Latcinnik and Berant (2020) proposed a method to generate free-text explanations supervised only on task signal, and Brahman et al. (2021) used sources of weak supervision to generate explanations for defeasible inference. Paranjape et al. (2021) design hand-crafted templates which they use with mask-infilling to produce contrastive explanations from pretrained language models. Concurrent work (Marasović et al., 2021) also investigates prompting; they study the effects of prompt format and model size on explanation quality. In contrast, we investigate generated explanations through fine-grained crowdsourcing evaluations, study the effect of prompt quality, and investigate a filtration method trained on human acceptability judgements.

Supervising on human preferences. Prior work has used binary judgements from crowdworkers to fit models to human preferences for summarization (Ziegler et al., 2019; Stiennon et al., 2020). West et al. (2021) demonstrate that GPT-3 + a supervised acceptability filter can generate a high-quality causal knowledge graph: in addition to their work being conducted in a different domain, our success conditions and evaluation metrics differ because we must produce a prediction for each instance (whereas they can simply discard bad generations).

6 Conclusion

We demonstrate GPT-3’s capacity to generate free-text explanations for NLP task instances in a few-shot setting. We further improve this capability via an overgenerate + filter approach, where the filter is trained on supervision from human acceptability ratings. We hope our results can guide future work on free-text explanations via neural or neuro-symbolic systems (Brahman et al., 2021; Majumder et al., 2021; Saha et al., 2021).

While human rationales for decision making are not necessarily the same as model rationales, the goal behind modeling human acceptability is to build trust with a human user. This trust may or may not be warranted (Jacovi et al., 2021); future work would be well-suited to further investigate generated explanations for incorrect label predictions, which could mislead end users.
References


of the International Conference on Computer Vision (ICCV).


Let's explain classification decisions.

Table 8: Example of a prompt with 3 training examples for SNLI: presented are the premise/hypothesis pairs, the gold labels, and the explanations (written by us) that act as input to GPT-3 (in practice, we use 8-24 examples per prompt). The text generated by the model acts as the free-text explanation. In this case, the model greedily auto-completes (given 12 examples): “A dog cannot carry something while asleep”.

A Prompt Construction

Following Perez et al. (2021), we avoid prompt tuning on the full training and development sets of the datasets studied, in order to ensure that our methods represent a true few-shot setting. To develop the initial prompt design, we experimented with no more than 10 different layouts in the GPT-3 Sandbox platform using 15 training examples on the CoS-E and e-SNLI datasets. For subsequent prompt design, we again used these 15 training examples for each dataset from which we sampled 6 prompts, along with a fixed 100-example “development set” randomly drawn from the training set. We preserve the “few-shot” approach by using a maximum of these same 115 instances to develop our prompting methods. For these 115 examples, the authors of this paper manually wrote high-quality explanations to be used as prompt examples (Table 10). As presented in Table 8, we found that structuring SNLI as a question-answering task achieved the best performance, similarly to Zhao et al. (2021). We provide an example of our SNLI prompt in Table 8 and CommonsenseQA in Table 9.

Let’s explain classification decisions.

Table 9: Example of a prompt with 3 training examples for CommonsenseQA: presented are the question and answer choices, the gold labels, and the explanations (written by us) that act as input to GPT-3 (in practice, we use 8-24 examples per prompt). The text generated by the model acts as the free-text explanation. In this case, the model greedily auto-completes (given 8 examples): “After peeing, the bladder is empty.”

In-context learning methods have been shown to have high variance based on hyperparameters including example order, number of examples given, and which examples are given (Jiang et al., 2020; Zhao et al., 2021; Lu et al., 2021). While these values have not been standardized, two prominent papers, Schick and Schütze (2021b) and Brown et al. (2020), use 32 and 64 prompt examples, respectively. Due to the 2049-token limit of the OpenAI GPT-3 API and the fact that the addition of explanations elongates each prompt instance, we find the maximum number of examples the API can accommodate is 24 for CoS-E, e-SNLI, and our handwritten explanations and 16 for ECQA. The focus of this work is not on finding the optimal prompt, but on developing a general strategy for few-shot explanation generation that could be
successful when no additional (large) validation set for tuning is available. Therefore, to provide as robust of an expected performance estimate as possible, we do not tune the additional hyperparameters, instead sampling them to approximate performance. Namely, while prior work uses one fixed prompt for all instances and varies the random seed, we approximate the same expected performance by sampling a new set of prompts for each instance. We also sample the number of prompts for each instance (and shuffle their order) from the values \{8, 16, 24\} for CommonsenseQA experiments, \{8, 16\} for experiments using ECQA explanations, and \{12, 18, 24\} for SNLI experiments (to maintain label balance). To overcome label bias in prompt ordering, for tasks with distinct answer choices per instance (CommonsenseQA), we shuffle the answer choices. For tasks with fixed answer choices (SNLI), we sample an equal number of prompt instances for each label (so number of prompt instances is a multiple of 3).

Table 10 shows a few non-cherry-picked examples of our handwritten explanations used as prompts relative to the datasets.

B Crowdsourcing Details

B.1 Head-to-Head Interface Details

We show the user the task input and gold label, and ask them to select which of two explanations best explains the answer. We instruct workers to consider the gold label to be correct even if they disagree with it (CommonsenseQA instances can be subjective) and to ignore minor grammar and spelling mistakes such as improper upper-casing. Figures 4 and 5 show the evaluation interface.

B.2 Absolute Interface Details

Figures 6 and 7 show the absolute evaluation interface (minus the acceptability attribute, which is collected in a separate run of the study). Our interface is inspired by prior work from psychology and the social sciences (Leake, 1991; Gopnik, 1998; Lombrero, 2007; Zemla et al., 2017; Chiyah Garcia et al., 2018; Clinciu et al., 2021; Sulik et al., 2021). We iterated over 3-4 versions of the questions and UI design until we had optimized agreement rates as much as possible. Our resulting two-part evaluation consists of 7 questions:

---

Table 10: Examples of explanations used as prompts from various sources, including our handwritten explanations. Correct answers for CommonsenseQA are underlined.

<table>
<thead>
<tr>
<th>CommonsenseQA (Talmor et al., 2019)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Question:</strong> A cat can’t talk, but a cat can what?</td>
<td><strong>Answer choices:</strong> sleep all day, meow, shed fur, see king, live many years</td>
<td><strong>Our Explanation:</strong> A cat can meow as a way to vocalize.</td>
<td><strong>CoS-E Explanation:</strong> The cat is a small carnivorous mammal</td>
</tr>
<tr>
<td><strong>Label:</strong> entailment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ECQA Explanation:</strong> A cat can meow but cannot see the king. Meowing is how a cat communicates and not by sleeping all day, shedding fur or by living many years.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Question:</strong> “There are 10 apples on an apple tree. Three fall off. Now there are X apples.” What is this an example of?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Answer choices:</strong> park, coloring book, garden center, math problem, gravity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Our Explanation:</strong> A math problem is usually posed as a question that requires some operation such as subtraction or addition to answer.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CoS-E Explanation:</strong> webmath is designed to help you solve</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ECQA Explanation:</strong> Math problem is an arithmetical problem of addition, subtraction, multiplication or division. So “There are 10 apples on an apple tree. Three fall off. Now there are X apples.” is a math problem. All the other options aren’t problems to be examples of the given question.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Label:</strong> neutral</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SNLI (Bowman et al., 2015)

| **Premise:** A person on a horse jumps over a broken down airplane. | **Hypothesis:** A person is training his horse for a competition. | **Label:** neutral |
| **Our Explanation:** While it is possible that jumping a horse over an obstacle is part of a training routine for a competition, it is also possible that the horse ride is being done for pleasure, not necessarily for a competition (sp). | **e-SNLI Explanation:** the person is not necessarily training his horse |
| **Premise:** Children smiling and waving at camera | **Label:** entailment |
| **Hypothesis:** There are children present | | |
| **Our Explanation:** Since the children are part of the event of smiling at the camera, they are present at the event under discussion. | **e-SNLI Explanation:** The children must be present to see them smiling and waving. |

**Table 10:** Examples of explanations used as prompts from various sources, including our handwritten explanations. Correct answers for CommonsenseQA are underlined.

**Part 1: Context-Independent Evaluation** We first assess the explanation in isolation, i.e., these questions are presented to the user without revealing the question/context that the explanation is attempting to address:

1. *How factual is this statement?* (generally false,
3. **How much information does the explanation have to justify the answer?** (not enough, enough, or too much) This question is designed to test the extent to which the provided novel information is *adequate* or *sufficient* (Lei et al., 2016; Ehsan et al., 2019).13

4. **Is the explanation acceptable?** (yes or no) The final question is designed to assess annotators’ overall judgement of the explanation as a whole.

We only ask Question 2 if the answer to Question 1 is “yes” and Question 3 if the answer to Question 2 is yes, because they regard the new facts, information, or reasoning. We found that most prior work tends to lump added-value, relevance, and adequacy judgements into one “informativeness” judgement (Clinciu et al., 2021), which we felt was too course to allow for meaningful error analysis.

### B.3 Acceptability Interface Details

Figures 8 and 9 show the binary acceptability interface used to collect training and test data for the overgeneration filter model.

### B.4 Quality Control and Payment

We use Amazon Mechanical Turk (AMT), and calculate pay on a rate of $15/hour. Every few batches, we check to ensure that the median time taken per-annotator amounts to approximately this pay rate. While annotators do tend to speed up the more HITs we released, first-round median times were approximately 30 seconds per head-to-head evaluation HIT (thus paid at $0.12 each), 1 minute per absolute evaluation HIT (thus paid at $0.25 each), and 35-39 seconds per acceptability HIT (5 explanations; paid at $0.20 each).

We require annotators to be located in either Australia, Canada, New Zealand, the United Kingdom, or the United States, as a proxy for English competency.14 We require a past HIT approval rate of >98% and >5000 HITs approved. We do not allow annotators to participate who were previously on a block list from our past AMT studies.

---

12This decision is inspired by prior work in psychology, which finds that explanations are only good “to the extent that people find [them] satisfying” (Gopnik, 1998; Sulik et al., 2021).

13In practice, we do not find Turkers use the “too much information” option often, except in the case of ECQA dataset explanations. We included the option because succinctness is an oft-cited explanatory virtue (Lombrozo, 2007; Zemla et al., 2015; Chiyah Garcia et al., 2018).

14We realize this is a broad assumption and likely sub-optimal. However, colleagues have found that broadening the geographical requirements often still leads to >90% of annotators in the US or Canada, due to AMT’s pay structure being optimal in these countries.
Annotators must complete a qualifying exam in order to participate in the main annotation tasks. The qualifying exam consists of 3 HITs in the same format as the main absolute evaluation task for CommonsenseQA. We pay $2.25 for the qualifying exam. There are 9-18 questions in total (3-6 questions per HIT), some of which are only answerable conditioned on previous answers. A user who answers “no” to question 3, for example, will not be asked to answer questions 4 and 5. Given the challenging and sometimes ambiguous nature of some of the questions, for the first run of the qualification exam, we manually awarded qualifications by inspecting the annotators’ answers. Scores for the first run compared to our answers (out of 17 annotators attempting) ranged from 5 to 14 out of 18. The median accuracy was 11 out of 18, and we find that awarding the qualification to those with scores at or above the median aligns closely with our manual inspection. We thus use this score to assign qualifications in future iterations.

Because it is necessary that annotators understand the task before they can evaluate quality (Wiegrefe and Marasović, 2021), for tasks that are more difficult, i.e., NLI, we additionally require annotators to pass (score of 7 or above) a task-specific qualification exam with 8 questions, paid at $1.25.

In order to track quality throughout evaluation, we compute inter-annotator agreement using Krippendorff’s $\alpha$ and use a hidden built-in Javascript function to compute time per HIT spent. If any annotator completed the tasks in an unreasonably low time, or removing their annotations substantially improves Krippendorff’s $\alpha$, we remove their annotations and re-annotate their instances. We additionally ensure that each experiment has a substantial number of distinct crowdworkers to mitigate individual annotator bias, reporting this as well as the mean and median number of HITs completed by each in Table 15.

### B.5 Statistics

The number of distinct crowd annotators and the median and mean number of HITs completed for each experiment can be found in Table 15. More detailed breakdowns of inter-annotator agreement for some experiments are in Tables 13 and 14.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Split</th>
<th># Instances by Agreement</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0/3</td>
<td>1/3</td>
</tr>
<tr>
<td>Com.QA</td>
<td>Train</td>
<td>932</td>
<td>1078</td>
</tr>
<tr>
<td></td>
<td>Dev</td>
<td>105</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>298</td>
<td>227</td>
</tr>
<tr>
<td>SNLI</td>
<td>Train</td>
<td>2372</td>
<td>805</td>
</tr>
<tr>
<td></td>
<td>Dev</td>
<td>272</td>
<td>87</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>678</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>Test2</td>
<td>666</td>
<td>234</td>
</tr>
</tbody>
</table>

Table 11: Statistics of our acceptability annotations.

### C 2/3 Acceptability Statistics

When we treat explanations rated by at least 2/3 annotators as “acceptable”, for CommonsenseQA, 77.9% of greedy-decoded explanations are acceptable; for SNLI, 51.0%. 50.5% of sampled explanations are acceptable; for SNLI, 23.5%. Out of the set of 5 (1 greedy + 4 stochastic), 97.7% of CommonsenseQA instances have at least one acceptable explanation, and 79.5% of SNLI.

### D Filter Model Details

We split the 4,955 distinct annotated explanations for CommonsenseQA (5000 for SNLI) into a train/dev set of 4500/455 (4500/500 for SNLI), where all 5 explanations for a given instance are placed in the same set to avoid leakage. We present statistics on the label distribution in Table 11. Along with the metric settings reported in the paper (“select-1” and explanation-level), we computed a metric that is instance-level but considers all explanations by computing metrics over the 5 explanations of an instance and then averaging across instances, finding in practice that the results are highly similar to the explanation-level evaluation.

We use Huggingface Datasets (Lhoest et al., 2021) and Huggingface Transformers (Wolf et al., 2020) for implementation. The T5-Large model is trained using a learning rate of $1E-4$ with linear decay, a batch size of 64, and default values for Adam (Kingma and Ba, 2015), gradient clipping, and dropout. We train for a maximum 200 epochs, performing early stopping on the validation loss with a patience of 10 epochs.

For T0-3B, we train with a batch size of 50. We use AdaFactor (Shazeer and Stern, 2018) with a linear warmup of 500 steps. We conduct an initial hyperparameter sweep over learning rate, considering $1E-5, 5E-05, 5E-06$. The learning
### SNLI (Bowman et al., 2015)

**Premise:** There are two kilted men, one of them older and is holding bagpipes with the other one with a drum.  
**Hypothesis:** Two kiled (sp) men hold bagpipes  
**Label:** contradiction  
**NLL-Predicted Explanation:** The two kilted men are not holding bagpipes, they are holding a baggage and a drum.  
**Filter-Predicted Explanation:** Just because there are kilted men does not necessarily mean that they are holding bagpipes. This can be seen from the older kilted man is holding the bagpipes rather than the kilder (sp) one.

### CommonsenseQA (Talmor et al., 2019)

**Question:** The hardcovers were especially tall, so he removed a shelf on the what to make room?  
**Answer choices:** hold alcohol, grocery store, bookcase, nightstand, or chest of drawers  
**NLL-Predicted Explanation:** Bookcases are tall, and so if they are full of hardcover books, they might need to be removed to make room for other things.  
**Filter-Predicted Explanation:** People keep books on bookshelves, and bookcases are very commonly used to hold books. Since hardcover books can be pretty tall, it takes up extra vertical space which would require removal of an upper shelf of the bookcase.

Table 12: Randomly-selected instances that our filter model predicted differently than NLL at the “select-1” task, and got incorrect but NLL got correct.

---

rate that achieves the best validation loss for the full-information model and the explanation-only model is $1E-5$, which we use for all training experiments. We format inputs to the model as follows:

```python
if explanation_only:
    input_string = (f"explanation: {expl}. Is this explanation good or bad?"
else:
    input_string = ('
{question} answer: {gold_label}.
"explanation: {expl}. "
Is this explanation good or bad?"
```

For CommonsenseQA, question, expl, and gold_label are: the commonsense QA question (with answer options omitted), the explanation candidate from GPT-3, and the true answer among the 5 options, respectively; for SNLI, premise: ... hypothesis: ..., the explanation candidate from GPT-3, and one of entailment/contradiction/neutral, respectively.

### Additional Filter Results

In the main experiments, at evaluation time, we labelled an explanation as acceptable if 3/3 annotators agreed on it. Here, we report results if this threshold is relaxed to 2/3. Overall, the results are comparable: T0-3B outperforms the baselines according to both select-1 accuracy and AP (see Table 18 and Table 19).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Split</th>
<th>Krippendorff’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonsenseQA</td>
<td>Training + Validation</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0.40</td>
</tr>
<tr>
<td>SNLI</td>
<td>Training + Validation</td>
<td>0.51</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Test2</td>
<td>0.47</td>
</tr>
</tbody>
</table>

Table 13: Inter-annotator agreement for acceptability AMT studies.
Table 14: Inter-annotator agreement for absolute-comparison AMT studies, using Krippendorff’s $\alpha$.

Table 15: Total # of annotators and median # HITs completed per-annotator for each AMT study (out of 750 total # HITs unless otherwise specified = 3 annotators for each of 250 test instances).

Table 16: Statistics from the graphs plotted in Figure 2. Mean ± standard error presented; numbers in parenthesis indicate the number of datapoints, if not 250. *For SNLI, we modified the evaluation framework such that “Supports Label” was always answered instead of being conditioned on “New Info”. Statistical significance results using a one-sided Wilcoxon signed-rank test at $p$-values of $\frac{1}{2} = 0.00001$, $\frac{1}{2} = 0.0001$, $\times = 0.01$, $\land = 0.1$ indicates that the median difference between the marked score distribution and the unmarked score distribution is greater than 0.
Table 18: Results for acceptability classifiers trained on CommonsenseQA, with “acceptability” defined as: “at least 2/3 annotators labelled as acceptable.” Subscripts indicate standard error over models trained with 5 different random seeds.

<table>
<thead>
<tr>
<th>Set of Test Explanations</th>
<th>Generality</th>
<th>Factuality</th>
<th>Grammar</th>
<th>New Info</th>
<th>Supports Label</th>
<th>Amount Info</th>
<th>Acceptability</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-3 Greedy for Com-QA</td>
<td>0.9 ± 0.1 (156)</td>
<td>0.8 ± 0.4 (153)</td>
<td>1.0 ± 0.1 (156)</td>
<td>0.1 ± 0.6 (156)</td>
<td>0.3 ± 0.7 (156)</td>
<td>0.5 ± 0.5 (117)</td>
<td>0.8 ± 0.7 (156)</td>
</tr>
<tr>
<td>GPT-3 Filtered for Com-QA</td>
<td>0.9 ± 0.3 (156)</td>
<td>0.7 ± 0.3 (155)</td>
<td>0.8 ± 0.4 (156)</td>
<td>0.7 ± 0.4 (156)</td>
<td>0.9 ± 0.4 (154)</td>
<td>0.2 ± 0.3 (152)</td>
<td>0.6 ± 0.6 (156)</td>
</tr>
<tr>
<td>GPT-3 Greedy for SNLI</td>
<td>0.8 ± 0.4 (91)</td>
<td>0.6 ± 0.6 (91)</td>
<td>0.9 ± 0.3 (91)</td>
<td>0.0 ± 0.7 (91)</td>
<td>-0.2 ± 0.6 (91)</td>
<td>-0.2 ± 0.5 (91)</td>
<td>-0.5 ± 0.7 (91)</td>
</tr>
<tr>
<td>GPT-3 Filtered for SNLI</td>
<td>0.8 ± 0.5 (91)</td>
<td>0.7 ± 0.4 (88)</td>
<td>0.7 ± 0.4 (91)</td>
<td>0.5 ± 0.6 (91)</td>
<td>0.5 ± 0.5 (91)</td>
<td>0.0 ± 0.3 (89)</td>
<td>0.1 ± 0.8 (91)</td>
</tr>
</tbody>
</table>

Table 17: Statistics from the graphs plotted in Figure 3. See the caption of Table 16 for further details.

<table>
<thead>
<tr>
<th>Model/Split →</th>
<th>“Select-1” Acc@2/3</th>
<th>Explanation-level AP@2/3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
<td>Test</td>
</tr>
<tr>
<td>Random</td>
<td>57.5±0.4</td>
<td>57.9±0.4</td>
</tr>
<tr>
<td>Constant</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>NLL</td>
<td>79.1</td>
<td>79.6</td>
</tr>
<tr>
<td>T0-3B Expl.-only</td>
<td>77.1±3.5</td>
<td>75.8±1.2</td>
</tr>
<tr>
<td>T0-3B</td>
<td>86.6±0.9</td>
<td>85.8±0.7</td>
</tr>
<tr>
<td>Oracle Upper-Bound</td>
<td>97.8</td>
<td>97.6</td>
</tr>
</tbody>
</table>

Table 19: Results for acceptability classifiers trained on SNLI with “acceptability” defined as: “at least 2/3 annotators labelled as acceptable.” Subscripts indicate standard error over models trained with 5 different random seeds.
Figure 4: An overview of the user interface of our head-to-head comparison AMT studies for CommonsenseQA. The top shows the instructions and the bottom the actual task. The Examples tab is collapsed here; shown in full in Figure 5.
Figure 5: The Examples tab given in the user interface of our head-to-head comparison AMT studies for Common-senseQA. The full interface is shown in Figure 4.
Figure 6: An overview of the user interface template of our absolute comparison AMT studies for Common-senseQA. The top shows the instructions and the bottom the actual task. Only part 1 of the task is shown here (part 2 appears once part 1 is submitted). The Main Example and More Examples tabs illustrating both parts 1 and 2 are collapsed here; see Figure 7.
Figure 7: The Main Example given in the user interface template of our absolute comparison AMT studies for CommonsenseQA. This format follows the actual task layout. The full interface is shown in Figure 6.
Figure 8: An overview of the user interface of our explanation acceptability AMT studies for CommonsenseQA. The top shows the instructions and the bottom the actual task. The "examples" tab is collapsed here; shown in full in Figure 9.
| Question: A bald eagle flies over St. Paul, where is it?  
Answer: minnesota |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Which explanations are acceptable? Select all that you believe adequately explain the answer (this may be none).</td>
</tr>
<tr>
<td>Bald eagles are native to the United States, but are rarer in Minnesota than in other states because Minnesota has a small supply of their primary foods (fish, mice, etc.).</td>
</tr>
<tr>
<td>St. Paul is a city in Minnesota, so a bald eagle flying over it would likely live in Minnesota.</td>
</tr>
<tr>
<td>St. Paul is in Minnesota.</td>
</tr>
<tr>
<td>The question assumes that you've seen the bald eagles that live in Minnesota.</td>
</tr>
<tr>
<td>Bald eagles live in northern North America, and can be found anywhere there. Given that one is flying over St. Paul, it must be either migrating, or it could be a photographer, making its appearance in the city unusual.</td>
</tr>
</tbody>
</table>

**Possible answer:**
Bald eagles are native to the United States, but are rarer in Minnesota than in other states because Minnesota has a small supply of their primary foods (fish, mice, etc.).

St. Paul is a city in Minnesota, so a bald eagle flying over it would likely live in Minnesota.

St. Paul is in Minnesota.

The question assumes that you've seen the bald eagles that live in Minnesota.

Bald eagles live in northern North America, and can be found anywhere there. Given that one is flying over St. Paul, it must be either migrating, or it could be a photographer, making its appearance in the city unusual.

One possible answer is that only the second and third explanations are acceptable. The first, fourth, and fifth explanations do not provide an explanation for the fact that St. Paul is in Minnesota.

Another possible answer is that only the third explanation is acceptable (since the second explanation incorrectly states that the bald eagle lives in Minnesota). However, we prefer that you err on the side of being forgiving if part of the explanation is correct.

---

| Question: Even if was clean and chlorinated, shallow water wouldn't make for much summer fun in your backyard what?  
Answer: swimming pool |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Which explanations are acceptable? Select all that you believe adequately explain the answer (this may be none).</td>
</tr>
<tr>
<td>It would be illegal to use and non-sanitary to use a lake or other body of water as a swimming pool.</td>
</tr>
<tr>
<td>Pools seem to be the kind of shallow water that people would swim in for fun.</td>
</tr>
<tr>
<td>Swimming pools get deeper than puddles or ponds.</td>
</tr>
<tr>
<td>Pools are the only body of water that contain both swimming-friendly water and shallow water.</td>
</tr>
<tr>
<td>Swimming pools are often used for summer fun.</td>
</tr>
</tbody>
</table>

**Possible answer:**
It would be illegal to use and non-sanitary to use a lake or other body of water as a swimming pool.

Pools seem to be the kind of shallow water that people would swim in for fun.

Swimming pools get deeper than puddles or ponds.

Pools are the only body of water that contain both swimming-friendly water and shallow water.

Swimming pools are often used for summer fun.

The question implies that the answer must be a body of water that can get deep, and must be able to be used for summer fun.

One possible answer is that only the fifth explanation is acceptable. The first, second, and fourth explanations do not explain that pools can get deep or that pools can be used for summer fun. The fifth explanation explains at least one of these facts.

Another possible answer is that both the third and the fifth explanations are acceptable, since the third explanation indirectly states that pools can get deep (although it is somewhat debatable if they are deeper than ponds).