Reframing Human-AI Collaboration for Generating Free-Text Explanations

Anonymous ACL submission

Abstract

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 crowd-sourced 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.

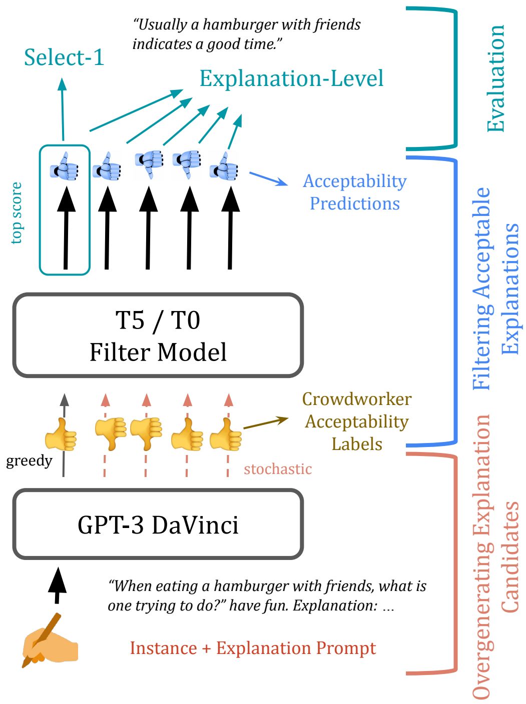


Figure 1: Illustration of our overgeneration + filtration pipeline for producing human-acceptable explanations for CommonsenseQA and SNLI (see examples in Table 1). Authors of this work write explanations to prompt GPT-3, generating 5 explanations per instance. An acceptability filter, trained on human binary acceptability judgments, determines which of these generated explanations are plausible. Evaluation is performed at both the explanation and the instance level.

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 crowdworks 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

[1] https://anonymous/

[2] We do not perform counterfactual human evaluations; ECQA explanations are thus often viewed by annotators as having "too much" information, see §2.3.

SNLI (Bowman et al., 2015)
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.

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.

internet. We prompt the model with several (question, answer and explanation)[3] triplets followed by an unexplained question-answer instance for which we expect the model to generate an explanation. We use a total of 115 randomly sampled train instances to create our prompts; each prompt consists of 8-24 randomly selected examples from this set. For each instance, we generate a single explanation with greedy decoding. More details about prompt construction are in Appendix A; example prompts are given in Tables 8-9.

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

[3]For simplicity, because we aim focus on free-text explanations (and not few-shot classification) we assume access to the gold label. In early experiments where we also had GPT-3 predict the label, we didn't observe significant differences in explanation quality between correct/incorrect instances.

	Preferred Explanation (%)			
Dataset	Gold-Standard	Tie	GPT-3	α
CoS-E	20.3	34.9	**44.8**	0.5
ECQA	**52.7**	12.9	34.4	0.2
e-SNLI	**63.6**	11.6	24.8	0.3

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 α. Results are shown as % preferences. GPT-3 explanations were prompted with explanations *from the respective datasets.*

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 α (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

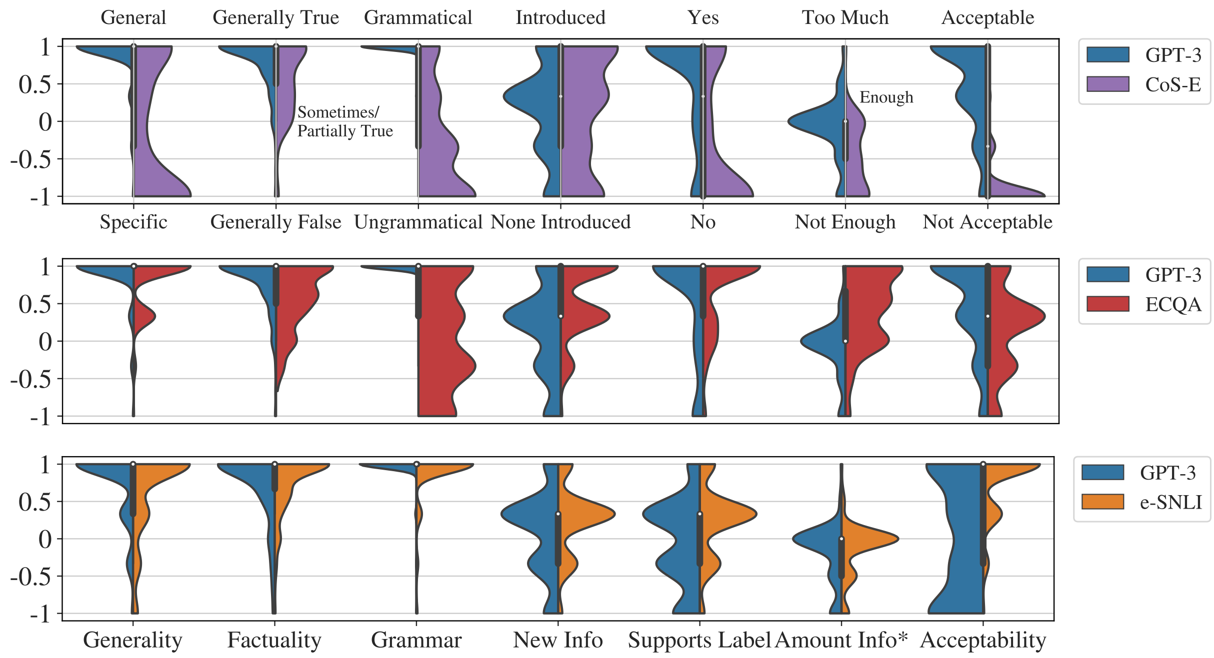


Figure 2: Absolute evaluation for GPT-3 and crowdsourced explanations for CommonsenseQA via CoS-E (top) and ECQA (middle) datasets, and for NLI via the e-SNLI dataset (bottom). The distribution of mean scores of 3 annotators for each instance in the test set is plotted. All attributes besides **Factuality** and **Amount Info** are binary. ***Amount Info** is the only attribute for which a value of 0 is preferred to a value of 1.

	Preferred GPT-3 Explanation (%)			
Dataset	**Gold Prompts**	**Tie**	**Our Prompts**	α
CoS-E	6.9	13.5	**79.6**	0.2
ECQA	15.9	9.5	**74.7**	0.3
e-SNLI	30.8	26.8	**42.4**	0.5

Table 3: Head-to-head human evaluation for 250 explanations generated by GPT-3 conditioned on either our handwritten explanations or gold crowdsourced explanations from the associated datasets.

	Preferred Explanation (%)			
Dataset	**Gold-Standard**	**Tie**	**GPT-3**	α
CoS-E	7.2	13.9	**78.9**	0.5
ECQA	44.5	9.7	**45.7**	0.4
e-SNLI	**49.6**	9.7	40.7	0.2

Table 4: Head-to-head human evaluation for 250 explanations generated by GPT-3 vs. written by crowdworkers in the datasets. GPT-3 explanations were prompted with explanations *handwritten by the authors*.

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 trough 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 criteria from §2.3, when considering GPT-3 explanations rated positively by 3/3 annotators as "acceptable", only 46.3% of our greedily decoded GPT-3-generated explanations for CommonsenseQA and 31.5% for NLI satisfy this criteria.[5]

Inspired by work in other generation tasks (Holtzman et al., 2020; Massarelli et al., 2020; Holtzman et al., 2021), we hypothesize that equally or more informative explanations can be generated by sampling stochastically. We sample 4 additional generations from GPT-3 for each instance, to complement the greedy generation. We again crowdsource 3 binary acceptability annotations for each new explanation.

While sampled explanations exhibit lower 3/3 acceptability than the greedy-decoded explanations (25.1% for CommonsenseQA; 11.3% for e-SNLI), their addition surprisingly results in a **much higher proportion** of instances that have *at least one* acceptable explanation in the set of 5. The greedy explanation was judged to be 3/3 acceptable in 46.3% of instances for CommonsenseQA and 31.5% for NLI; this increases to 79.5% and 51.2%, respectively, when sampled explanations are included.[6]

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

[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.

[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 not overfit to specfic annotators (Geva et al., 2019), we collect an additional set of judgments for the test set of SNLI 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.[7] 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.[8]

[7] Our results don't significantly change if a $2/3$ cutoff is used at test time instead: Appendix E contains the results.

[8] According to GPT-3, a sampled explanation has a lower NLL than greedy for only 2.8% and 3.6% of instances, respectively, of our CommonsenseQA and SNLI test sets.

	"Select-1" Acc@3/3		Expl.-level AP@3/3	
	Dev	Test	Dev	Test
Random	$26.8_{3.2}$	$30.6_{2.1}$	$27.4_{1.1}$	$31.6_{0.4}$
Constant	—	—	27.9	31.8
NLL	41.8	52.0	42.4	45.6
T5-L Expl.-only	$40.2_{3.9}$	$49.8_{1.1}$	$43.5_{1.5}$	$50.0_{1.9}$
T0-3B Expl.-only	$42.6_{1.4}$	$47.3_{1.6}$	$41.1_{2.0}$	$54.1_{1.7}$
T5-L w/o HA	$46.6_{2.3}$	$55.4_{3.2}$	$47.0_{2.3}$	$56.9_{3.6}$
T5-L	$46.4_{2.9}$	$55.4_{2.1}$	$45.1_{3.4}$	$58.3_{1.5}$
T0-3B w/o HA	$48.4_{2.0}$	$57.4_{2.8}$	$44.5_{2.3}$	$59.8_{2.7}$
T0-3B	$\mathbf{48.6_{0.9}}$	$\mathbf{59.9_{1.1}}$	$\mathbf{49.7_{1.6}}$	$\mathbf{64.0_{1.5}}$
Oracle U.B.	78.0	82.0	100.0	100.0

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 (§3).

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 (§2.3) 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-

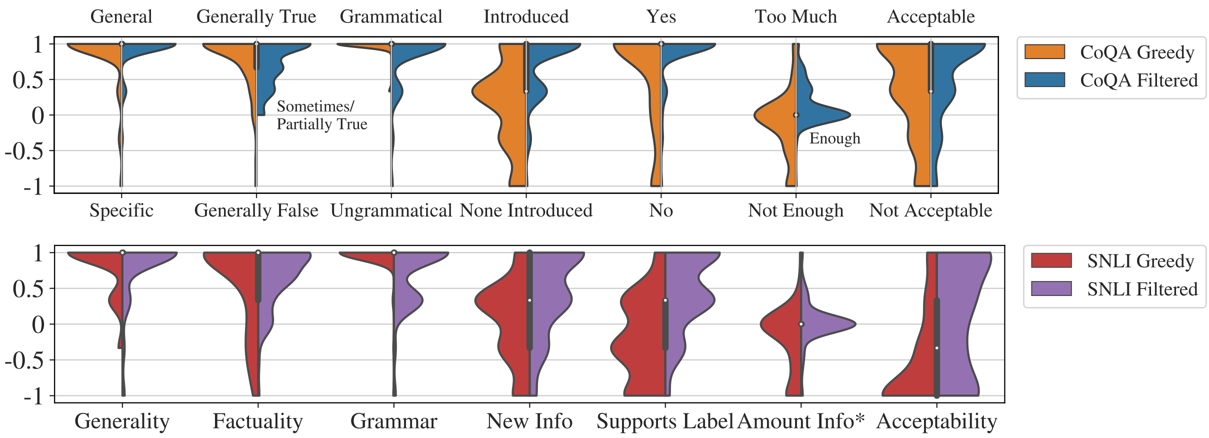


Figure 3: Absolute evaluation results in the "select-1" setting for the instances where our best-performing filter model does *not* select the greedy explanation (156 instances for CommonsenseQA (top); 91 for NLI (bottom)). See caption of Figure 2 and the Appendix-Table 17 for more details.

	"Select-1" Acc@3/3			Explanation-level AP@3/3		
	Dev	Test	Test2	Dev	Test	Test2
Random	$15.2_{0.7}$	$14.7_{0.1}$	$13.6_{0.3}$	$15.0_{0.6}$	$14.4_{0.3}$	$13.8_{0.2}$
Constant	—	—	—	15.2	14.5	13.7
NLL	33.0	32.0	31.6	29.9	32.7	28.5
Expl.-only	$30.2_{0.8}$	$30.9_{2.1}$	$27.8_{1.8}$	$30.6_{2.5}$	$30.6_{1.3}$	$25.9_{2.3}$
w/o HA	$\mathbf{38.2}_{\mathbf{1.9}}$	$38.5_{1.8}$	$\mathbf{36.2}_{\mathbf{1.4}}$	$\mathbf{49.3}_{\mathbf{5.3}}$	$\mathbf{48.5}_{\mathbf{3.3}}$	$\mathbf{52.8}_{\mathbf{5.3}}$
Full	$37.8_{0.8}$	$\mathbf{38.7}_{\mathbf{0.7}}$	$35.0_{1.2}$	$46.8_{3.6}$	$47.6_{3.5}$	$49.5_{4.8}$
Oracle U.B.	51.0	52.4	46.4	100.0	100.0	100.0

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

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 124 out of 250 instances for CommonsenseQA, and 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:[9] 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

[9]Because the input instances are identical in the select-1 setting between the 5 explanation candidates, any improvement over the explanation-only baseline is evidence of feature *interactions*, specifically.

SNLI
Premise: An officer in a black uniform and hat stands to the left of a large structure with other officers in the background. **Hypothesis:** An officer enjoys coffee in a shop. **Label:** contradiction **NLL-Predicted Explanation:** An officer in a black uniform and hat is not necessarily an officer enjoying coffee in a shop. **Filter-Predicted Explanation:** An officer in a structure standing to one side is not the same as enjoying coffee in a shop.
CommonsenseQA
Question: Where would there be an auditorium with only a single person speaking? **Answer choices:** theater, park, <u>university campus</u>, crowd, or lights **NLL-Predicted Explanation:** An auditorium is a large room used for lectures, and a single person speaking is likely to be a lecture. **Filter-Predicted Explanation:** On university or university-like campuses, the auditoriums are often used for lectures or other University events, where a single person might be speaking.

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.

factuality for NLI) are not significant.[10]

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.

[10] Krippendorff's α for these experiments is 0.32 for CommonsenseQA and 0.33 for SNLI.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3050–3065, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Faeze Brahman, Vered Shwartz, Rachel Rudinger, and Yejin Choi. 2021. Learning to rationalize for nonmonotonic reasoning with distant supervision. In *the AAAI Conference on Artificial Intelligence*.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Francisco Javier Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 99–108, Tilburg University, The Netherlands. Association for Computational Linguistics.

Miruna-Adriana Clinciu, Arash Eshghi, and Helen Hastie. 2021. A study of automatic metrics for the evaluation of natural language explanations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2376–2387, Online. Association for Computational Linguistics.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274. ACM.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Alison Gopnik. 1998. Explanation as orgasm. *Minds and machines*, 8(1):101–118.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision (ECCV)*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Ari Holtzman, Peter West, Vered Schwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. *arXiv preprint arXiv:2104.08315*.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings* *of the International Conference on Computer Vision (ICCV)*.

Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2288–2296.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. 2018. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Vivian Lai, Samuel Carton, and Chenhao Tan. 2020. Harnessing explanations to bridge ai and humans. In *Fair and Responsible AI Workshop at the CHI Conference on Human Factors in Computing Systems*.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. arXiv:2004.05569.

David B Leake. 1991. Goal-based explanation evaluation. *Cognitive Science*, 15(4):509–545.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

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

Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3):232–257.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. ArXiv preprint arXiv:2104.08786.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Rationale-inspired natural language explanations with commonsense. ArXiv preprint arXiv:2106.13876.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. 2021. Few-shot self-rationalization with natural language prompts. *arXiv preprint arXiv:2111.08284*.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! training text-to-text models to explain their predictions. arXiv:2004.14546.

Bhargavi Paranjape, Julian Michael, Marjan Ghazvininejad, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Prompting contrastive explanations for commonsense reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4179–4192, Online. Association for Computational Linguistics.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. ArXiv preprint arXiv:2105.11447.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI Blog.

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

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself!

leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. ArXiv:2110.08207.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Justin Sulik, Jeroen van Paridon, and Gary Lupyan. 2021. Explanations in the wild. PsyArXiv.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenhao Tan. 2021. On the diversity and limits of human explanations. ArXiv preprint arXiv:2106.11988.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. ArXiv preprint arXiv:2110.07178.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yordan Yordanov, Vid Kocijan, Thomas Lukasiewicz, and Oana-Maria Camburu. 2021. Few-shot out-of-domain transfer learning of natural language explanations. *arXiv preprint arXiv:2112.06204*.

Jeffrey C Zemla, Steven Sloman, Christos Bechlivanidis, and David A Lagnado. 2017. Evaluating everyday explanations. *Psychonomic bulletin & review*, 24(5):1488–1500.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the Thirty-Eighth International Conference on Machine Learning (ICML)*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. ArXiv preprint arXiv:1909.08593.

Let's explain classification decisions.
A young boy wearing a tank-top is climbing a tree.
question: A boy was showing off for a girl.
true, false, or neither? neither
why? A boy might climb a tree to show off for a girl, but he also might do it for fun or for other reasons.
###
A person on a horse jumps over a broken down airplane.
question: A person is outdoors, on a horse.
true, false, or neither? true
why? Horse riding is an activity almost always done outdoors. Additionally, a plane is a large object and is most likely to be found outdoors.
###
There is a red truck behind the horses.
question: The horses are becoming suspicious of my apples.
true, false, or neither? false
why? The presence of a red truck does not imply there are apples, nor does it imply the horses are suspicious.
###
A dog carries an object in the snow.
question: A dog is asleep in its dog house.
true, false, or neither? false
why?

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.
question: When remembering a tragedy in the past, what do many people feel?
pain, depression, knowing, knowledge, or nostalgia?
depression
why? Remembering a past tradedy can resurface feelings that arose in response to that tragedy. Because tragedies are not positive events, it's possible that sadness and depression could arise from remembering it.
###
question: What do people do sometimes when they find a good deal?
race cars, murder each other, believe in god, fight each other, or fear death?
fight each other
why? Malls sometimes have sales, e.g., on black friday, when they offer good deals; however, the items are sometimes in limited supply, which can cause altercations between folks, each trying to buy the same item.
###
question: What does someone who has a greed for energy do?
buy food, lie, get, cause businesses to grow, or win?
buy food
why? When consumed, food provides energy and satisfies the greed for it.
###
question: Immediately after peeing, a person's bladder is what?
full, empty, filled, stretchable, or collapsed?
empty
why?

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.[11] 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; Lombrozo, 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:

[11] Perez et al. (2021) show that performing cross-validation or tuning via maximum description length over a small validation set does not significantly outperform random sampling of these values.

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 competion (sp). **e-SNLI Explanation:** the person is not necessarily training his horse
Premise: Children smiling and waving at camera **Hypothesis:** There are children present **Label:** entailment **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.

CommonsenseQA (Talmor et al., 2019)
Question: A cat can't talk, but a cat can what? **Answer choices:** sleep all day, meow, shed fur, see king, live many years **Our Explanation:** A cat can meow as a way to vocalize. **CoS-E Explanation:** the cat is a small carnivorous mammal **ECQA Explanation:** 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.
Question: "There are 10 apples on an apple tree. Three fall off. Now there are X apples." What is this an example of? **Answer choices:** park, coloring book, garden center, math problem, gravity **Our Explanation:** A math problem is usually posed as a question that requires some operation such as subtraction or addition to answer. **CoS-E Explanation:** webmath is designed to help you solve **ECQA Explanation:** 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.

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,

sometimes or partially true, generally true, or need more information to judge). This question is designed to test both generality (can the explanation's truthfulness be ascertained or is more information needed?) and factuality, which aligns with "compatibility with receiver's existing beliefs" and that the best explanation is the "most likely" explanation from the receiver/user's perspective (Lombrozo, 2007; Zemla et al., 2017; Sulik et al., 2021). Generality is coded based on whether a truthfulness answer is selected (considered to be general) or whether the "need more information to judge" choice is selected (considered not to be general).

2. *Is this statement grammatical?* (yes or no) This question is designed to test for clarity, aligning with characteristics such as coherence (Lei et al., 2016) and human-likeness and understandability (Ehsan et al., 2019).

Part 2: Context-Dependent Evaluation We next show the user the question (premise and hypothesis for SNLI) and gold answer that the explanation was conditioned on. We then ask:

1. *Does the explanation provide new facts, information or reasoning not stated in the question and answer?* (yes or no) In our preliminary experiments, we found some explanations simply restate the question declaratively with the answer filled in. This question addresses the distinction between "validity" and "utility" (Leake, 1991): an explanation can be valid (i.e., a restatement of the question with the answer filled-in might be correct), but not useful; utility is defined by whether an explanation "satisfies an explainer's need for information". And while utility is best understood in the context of real-world applications (Lai et al., 2020), we nonetheless aim to identify vacuous explanations that do not provide new information.
2. *Is the new information relevant to justifying the answer? (yes or no)* New information, if provided, "should be compatible with our existing beliefs, and consistent with the evidence and with itself" (Zemla et al., 2017). This question is designed to test whether the information provided supports the label. The specific interpretation of "relevance" is purposefully left to the annotator.[12]
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* (Kim et al., 2016; 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.

[12]This 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).

[13]In 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., 2017; Chiyah Garcia et al., 2018).

[14]We realize this is a broad assumption and likely suboptimal. 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 explanation quality (Wiegreffe 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 α 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 α, 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.

Dataset	Split	# Instances by Agreement 0/3	1/3	2/3	3/3	Total
Com.QA	Train	932	1078	1194	1296	4500
	Dev	105	91	132	127	455
	Test	298	227	328	397	1250
SNLI	Train	2372	805	621	702	4500
	Dev	272	87	65	76	500
	Test	678	225	166	181	1250
	Test2	666	234	179	171	1250

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 bagpipe 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:

```
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.

E 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).

Dataset	Split	Krippendorff's α
CommonsenseQA	Training + Validation	0.32
	Test	0.40
SNLI	Training + Validation	0.51
	Test	0.50
	Test2	0.47

Table 13: Inter-annotator agreement for acceptability AMT studies.

AMT Study	Dataset	Generality	Factuality	Grammar	New Info	Supports Label	Amount Info	Acceptability	Aggregate
GPT-3 Greedy	Com.QA	0.37	0.32	-0.01	0.09	0.45	0.21	0.28	**0.38**
GPT-3 Greedy	SNLI	0.25	0.57	0.39	-0.01	0.04	0.17	0.52	**0.40**
Dataset	CoS-E	0.71	0.38	0.36	0.42	0.68	0.08	0.25	**0.59**
Dataset	ECQA	0.01	0.21	0.30	0.00	0.03	0.25	0.04	**0.20**
Dataset	e-SNLI	0.37	0.23	0.37	-0.14	-0.12	0.04	0.15	**0.19**
GPT-3 Filtered	Com.QA	0.25	0.18	0.25	0.11	0.27	0.15	0.30	**0.32**
GPT-3 Filtered	SNLI	0.41	0.19	0.07	0.17	0.08	0.13	0.50	**0.33**

Table 14: Inter-annotator agreement for absolute-comparison AMT studies, using Krippendorff's α computed on an interval scale from -1 to 1. The aggregate score is computed by treating the annotations along each attribute as separate instances and computing agreement across the entire set.

AMT Study	Task/Dataset	# Annotators	Median # HITs (Mean)
GPT-3 Greedy w/ Dataset Prompts vs. Dataset	Com.QA/CoS-E	16	31.5 (46.9)
	Com.QA/ECQA	13	35 (57.7)
	e-SNLI	12	39 (62.5)
GPT-3 Greedy: Author-written vs. Dataset Prompts	Com.QA/CoS-E	7	84 (107.1)
	Com.QA/ECQA	13	49 (57.7)
	e-SNLI	8	43.5 (93.8)
GPT-3 Greedy w/ Author-written Prompts vs. Dataset	Com.QA/CoS-E	8	90 (93.8)
	Com.QA/ECQA	17	27 (44.1)
	e-SNLI	8	93 (93.8)
GPT-3 Greedy (Absolute)	Com.QA	13	51 (57.7)
	SNLI	12	14 (62.5)
Dataset (Absolute)	CoS-E	14	58 (53.6)
	ECQA	19	7 (39.5)
	e-SNLI	13	16 (57.7)
Acceptability (Training and Validation Data)	Com.QA (2973 HITs)	34	70 (87.4)
	SNLI (3000 HITs)	14	128.5 (214.3)
Acceptability (Test Data)	Com.QA	17	32 (44.1)
	SNLI	11	26 (68.1)
	SNLI (Test2)	7	65 (107.1)
	CoS-E	13	48 (57.7)
	ECQA	16	38.5 (46.9)
	e-SNLI	9	60 (83.3)
GPT-3 Filtered (Absolute)	Com.QA (468 HITs)	10	44.5 (46.8)
	SNLI (273 HITs)	6	53 (45.5)

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

Set of Test Explanations	Generality	Factuality	Grammar	New Info	Supports Label	Amount Info	Acceptability
GPT-3 Greedy for Com.QA	$\mathbf{0.9 \pm 0.4}^{\ddagger}$	$\mathbf{0.8 \pm 0.4}$ (247)†	$\mathbf{1.0 \pm 0.1}^{\ddagger}$	0.1 ± 0.6	$\mathbf{0.5 \pm 0.7}$ (217)‡	$\mathbf{-0.1 \pm 0.4}$ (186)‡	$\mathbf{0.5 \pm 0.6}$‡
CoS-E	-0.2 ± 0.9	0.5 ± 0.5 (131)	-0.3 ± 0.7	0.1 ± 0.8	-0.3 ± 0.9 (190)	-0.5 ± 0.5 (78)	-0.9 ± 0.4
GPT-3 Greedy for Com.QA	$\mathbf{0.9 \pm 0.4}^{\vee}$	$\mathbf{0.8 \pm 0.4}^{\ddagger}$ (247)	$\mathbf{1.0 \pm 0.1}^{\ddagger}$	0.1 ± 0.6	0.5 ± 0.7 (217)	$\mathbf{-0.1 \pm 0.4}$ (186)	$\mathbf{0.5 \pm 0.6}$‡
ECQA	0.8 ± 0.4	0.6 ± 0.4 (249)	0.1 ± 0.7	$\mathbf{0.6 \pm 0.5}^{\ddagger}$	$\mathbf{0.7 \pm 0.5}$ (247)$^{\wedge}$	0.5 ± 0.5 (239)‡	0.1 ± 0.6
GPT-3 Greedy for SNLI	$\mathbf{0.7 \pm 0.5}^{\wedge}$	0.7 ± 0.5 (246)	$\mathbf{1.0 \pm 0.2}^{\dagger}$	0.1 ± 0.6	$-0.1 \pm 0.6^{*}$	-0.2 ± 0.4 (203)	0.0 ± 0.8
e-SNLI	0.6 ± 0.6	0.8 ± 0.4 (236)	0.9 ± 0.4	$\mathbf{0.2 \pm 0.5}^{\vee}$	$\mathbf{0.2 \pm 0.5}^{*\ddagger}$	$\mathbf{-0.1 \pm 0.4}$ (238)$^{\wedge}$	$\mathbf{0.7 \pm 0.4}^{\ddagger}$

Table 16: Statistics from the graphs plotted in Figure 2. Mean $\pm$ 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 $\ddagger = 0.00001$, $\dagger = 0.0001$, $\vee = 0.01$, $\wedge = 0.1$ indicates that the median difference between the marked score distribution and the unmarked score distribution is greater than 0.

Set of Test Explanations	Generality	Factuality	Grammar	New Info	Supports Label	Amount Info	Acceptability
GPT-3 **Greedy** for Com.QA	0.9 ± 0.4 (156)	$\mathbf{0.8 \pm 0.4}$ (153)$^{\vee}$	$\mathbf{1.0 \pm 0.1}$ (156)‡	0.1 ± 0.6 (156)	0.5 ± 0.7 (135)	$\mathbf{-0.1 \pm 0.5}$ (117)	0.3 ± 0.7 (156)
GPT-3 **Filtered** for Com.QA	$\mathbf{0.9 \pm 0.3}$ (156)$^{\wedge}$	0.7 ± 0.3 (155)	0.8 ± 0.4 (156)	$\mathbf{0.7 \pm 0.4}$ (156)‡	$\mathbf{0.9 \pm 0.4}$ (154)‡	0.2 ± 0.3 (152)‡	$\mathbf{0.6 \pm 0.6}$ (156)$^{\vee}$
GPT-3 **Greedy** for SNLI	0.8 ± 0.4 (91)	0.6 ± 0.6 (91)	$\mathbf{0.9 \pm 0.3}$ (91)‡	0.0 ± 0.6 (91)	$-0.2 \pm 0.6^{*}$ (91)	-0.2 ± 0.5 (66)	-0.5 ± 0.7 (91)
GPT-3 **Filtered** for SNLI	0.8 ± 0.5 (91)	0.7 ± 0.4 (88)	0.7 ± 0.4 (91)	$\mathbf{0.5 \pm 0.6}$ (91)‡	$\mathbf{0.5 \pm 0.5}^{*}$ (91)‡	$\mathbf{0.0 \pm 0.3}$ (89)†	$\mathbf{0.1 \pm 0.8}$ (91)‡

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

	"Select-1" Acc@2/3		Explanation-level AP@2/3	
↓Model/Split→	**Dev**	**Test**	**Dev**	**Test**
Random	$57.3_{0.4}$	$57.9_{0.4}$	$56.2_{0.5}$	$58.0_{0.9}$
Constant	—	—	56.9	58.0
NLL	79.1	79.6	77.5	75.0
T0-3B Expl.-only	$77.1_{3.5}$	$75.8_{1.2}$	$75.6_{2.0}$	$77.3_{1.4}$
T0-3B	$\mathbf{86.6_{0.9}}$	$\mathbf{85.8_{0.7}}$	$\mathbf{85.6_{0.5}}$	$\mathbf{87.0_{0.8}}$
Oracle Upper-Bound	97.8	97.6	100.0	100.0

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.

	"Select-1" Acc@2/3			Explanation-level AP@2/3		
↓Model/Split→	**Dev**	**Test**	**Test2**	**Dev**	**Test**	**Test2**
Random	$28.2_{0.5}$	$27.8_{0.2}$	$28.0_{0.1}$	$28.1_{0.9}$	$27.6_{0.3}$	$28.3_{0.6}$
Constant	—	—	—	28.2	27.8	28.0
NLL	51.0	51.2	50.4	47.7	47.5	46.1
T0-3B Expl.-only	$47.0_{1.0}$	$50.5_{2.1}$	$50.6_{2.8}$	$48.9_{1.4}$	$45.2_{1.5}$	$44.9_{2.1}$
T0-3B	$\mathbf{57.8_{1.9}}$	$\mathbf{60.3_{1.5}}$	$\mathbf{59.2_{2.3}}$	$\mathbf{66.7_{3.3}}$	$\mathbf{64.7_{3.3}}$	$\mathbf{67.1_{3.6}}$
Oracle Upper-Bound	76.0	81.2	77.6	100.0	100.0	100.0

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.

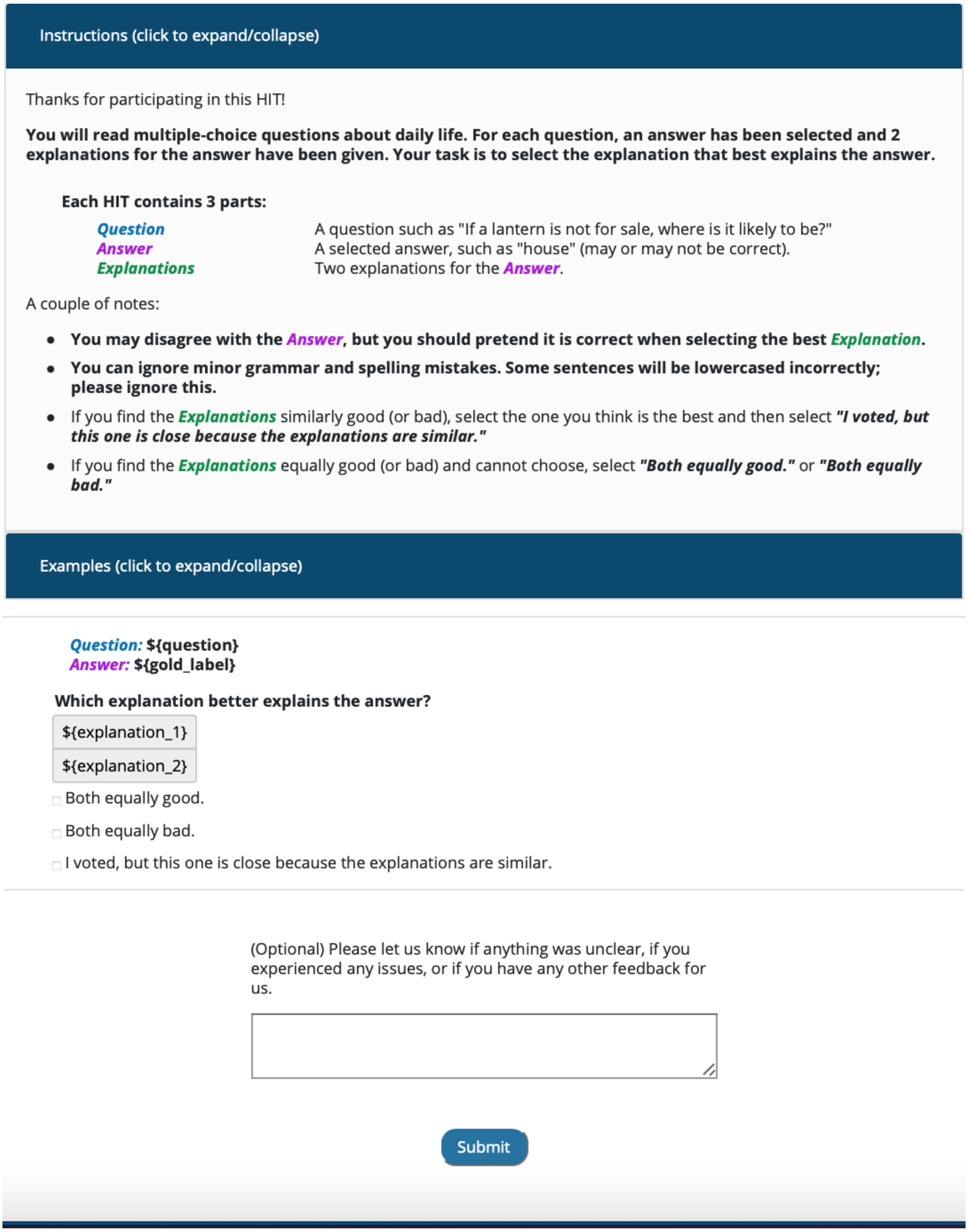


Figure 4: An overview of the user interface of our head-to-head comparison AMT studies for CommonsenseQA. The top shows the instructions and the bottom the actual task. The Examples tab is collapsed here; shown in full in Figure 5.

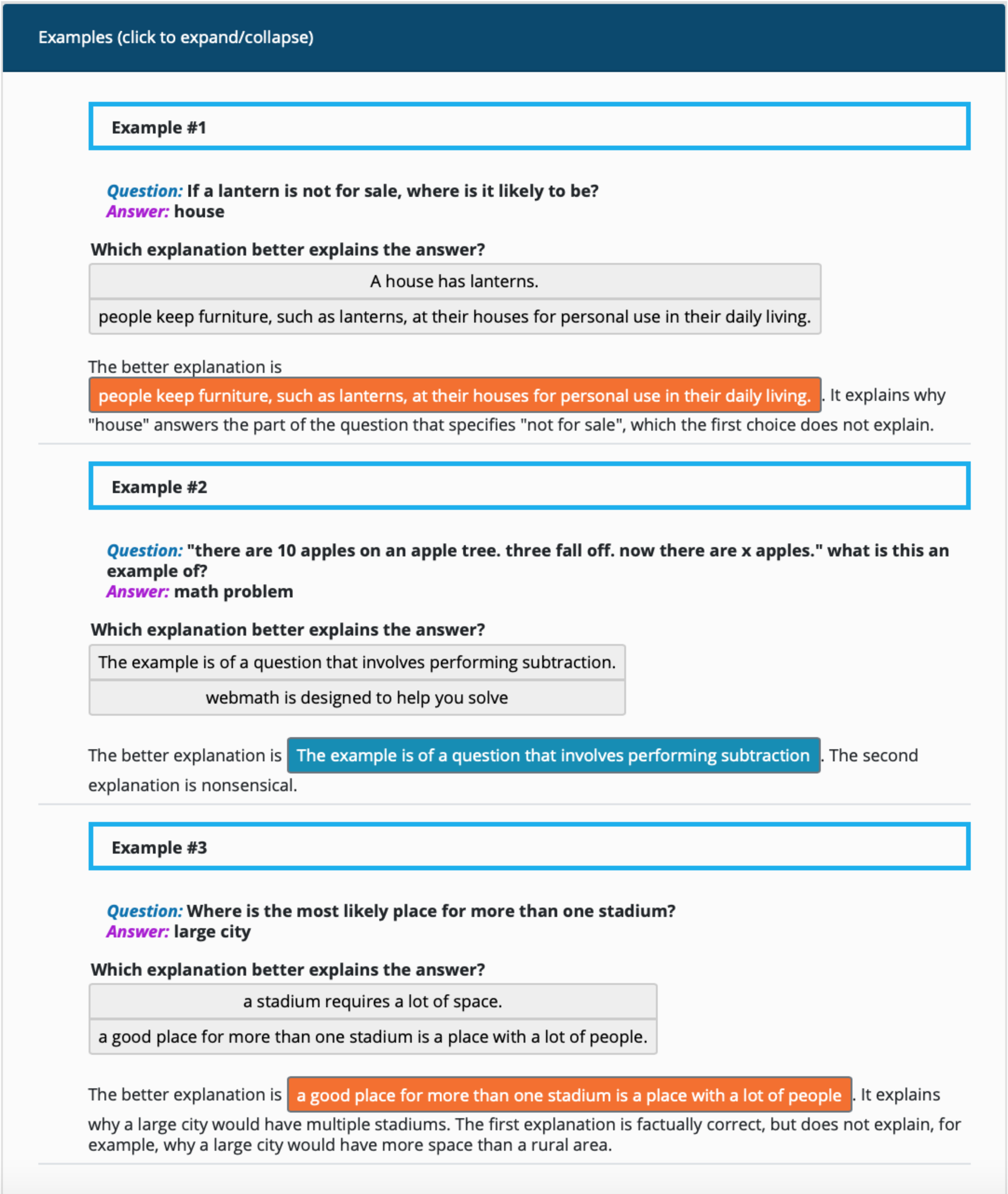


Figure 5: The Examples tab given in the user interface of our head-to-head comparison AMT studies for CommonsenseQA. The full interface is shown in Figure 4.

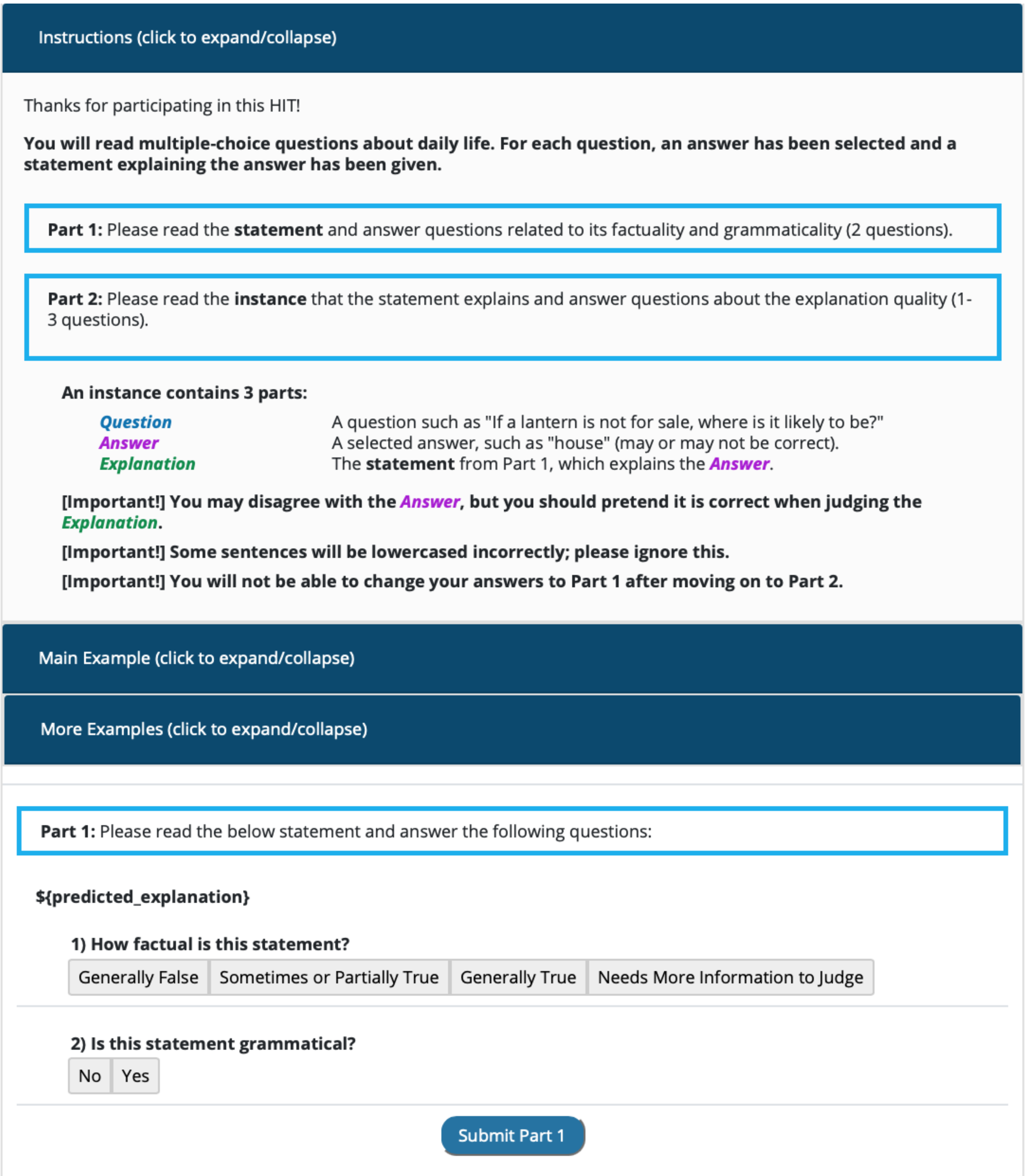


Figure 6: An overview of the user interface template of our absolute comparison AMT studies for CommonsenseQA. The top shows the instructions and the bottom the actual task. Only part 1 of the task is shown here (part 2 appears once part 1 is submitted). The Main Example and More Examples tabs illustrating both parts 1 and 2 are collapsed here; see Figure 7.

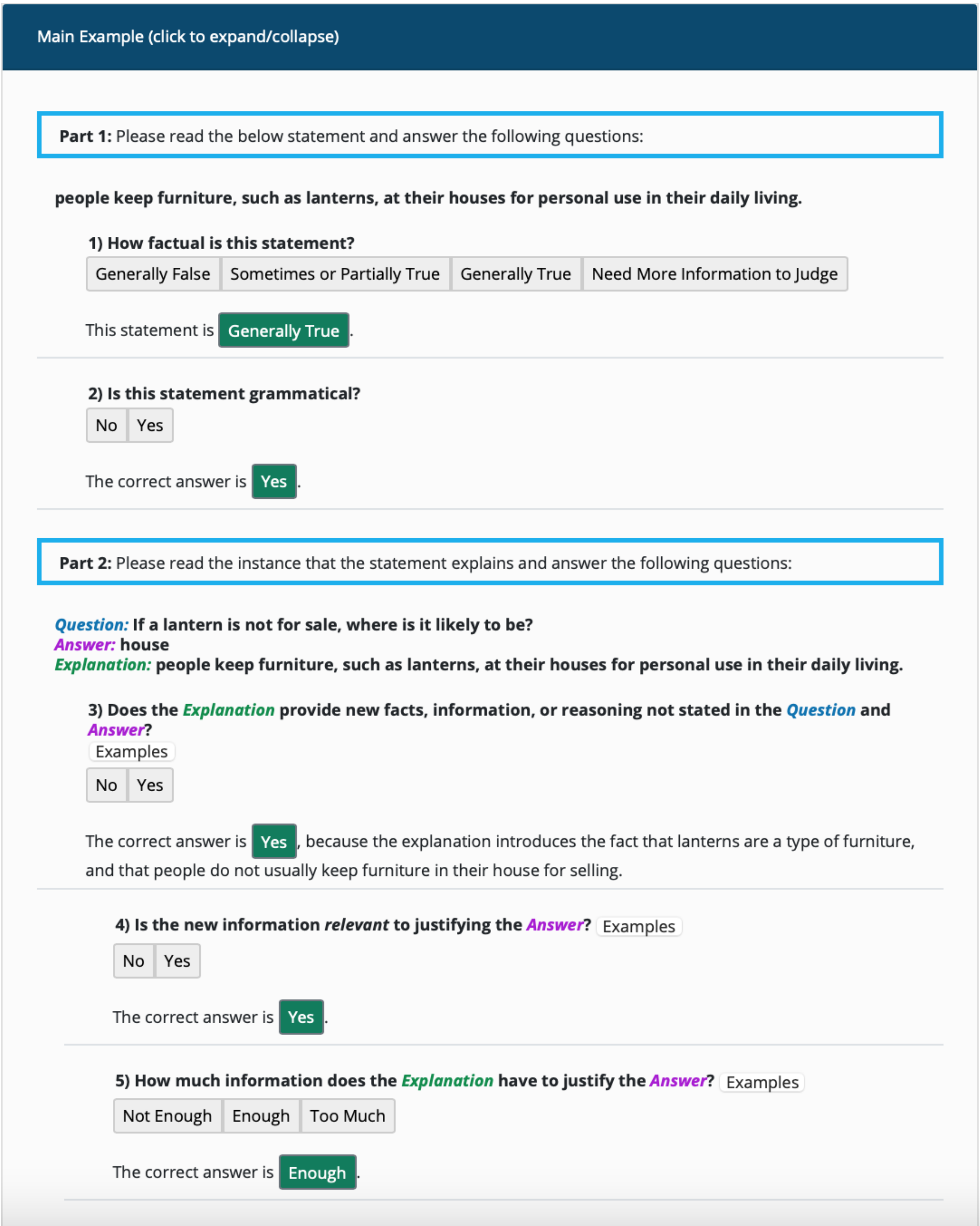
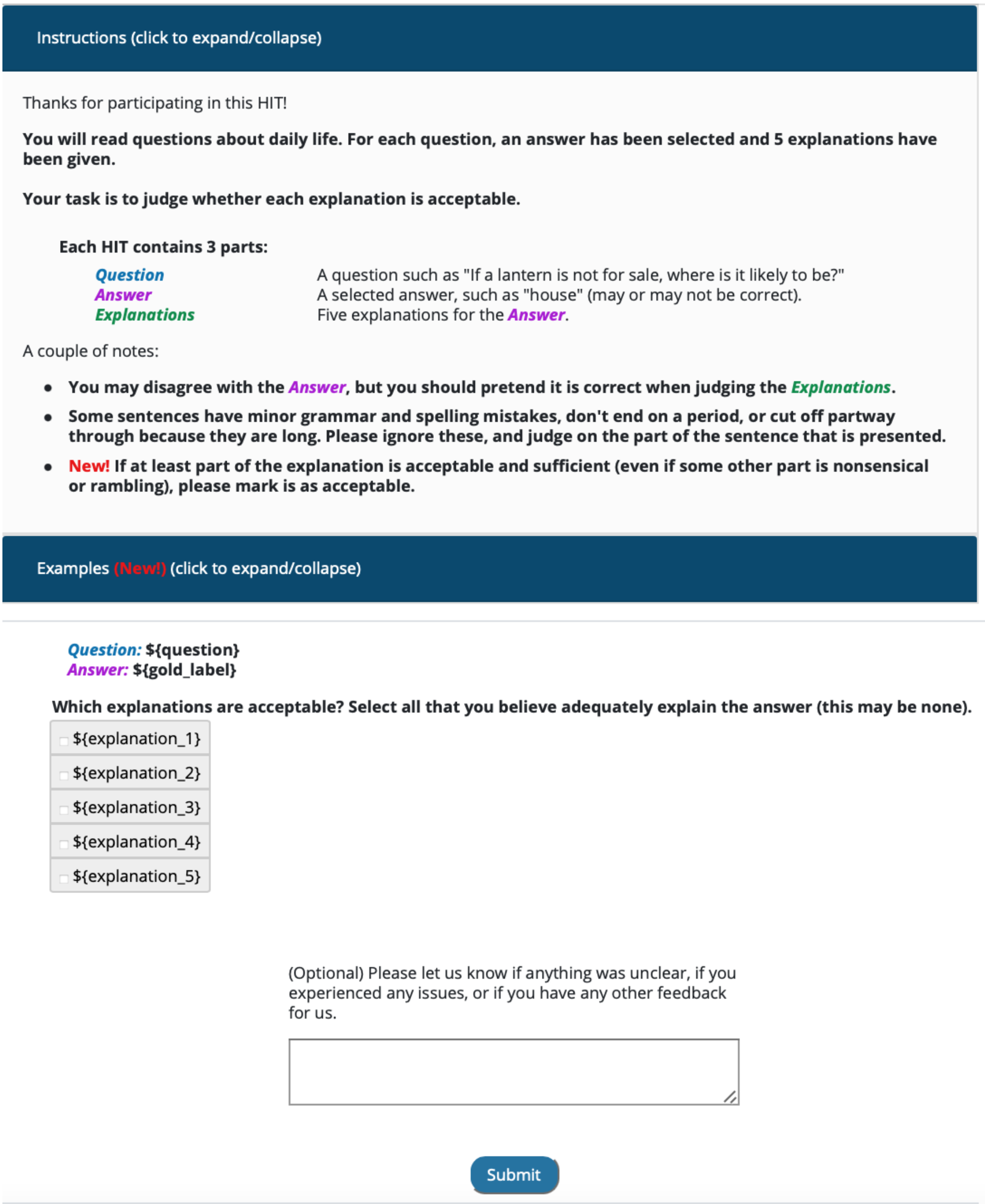


Figure 7: The Main Example given in the user interface template of our absolute comparison AMT studies for CommonsenseQA. This format follows the actual task layout. The full interface is shown in Figure 6.



Instructions (click to expand/collapse)

Thanks for participating in this HIT!

You will read questions about daily life. For each question, an answer has been selected and 5 explanations have been given.

Your task is to judge whether each explanation is acceptable.

Each HIT contains 3 parts:

Question	A question such as "If a lantern is not for sale, where is it likely to be?"
Answer	A selected answer, such as "house" (may or may not be correct).
Explanations	Five explanations for the ***Answer***.

A couple of notes:

- **You may disagree with the *Answer*, but you should pretend it is correct when judging the *Explanations*.**
- **Some sentences have minor grammar and spelling mistakes, don't end on a period, or cut off partway through because they are long. Please ignore these, and judge on the part of the sentence that is presented.**
- **New! If at least part of the explanation is acceptable and sufficient (even if some other part is nonsensical or rambling), please mark is as acceptable.**

Examples (New!) (click to expand/collapse)

Question: **${question}**
Answer: **${gold_label}**

Which explanations are acceptable? Select all that you believe adequately explain the answer (this may be none).

- ${explanation_1}
- ${explanation_2}
- ${explanation_3}
- ${explanation_4}
- ${explanation_5}

(Optional) Please let us know if anything was unclear, if you experienced any issues, or if you have any other feedback for us.

Submit

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.

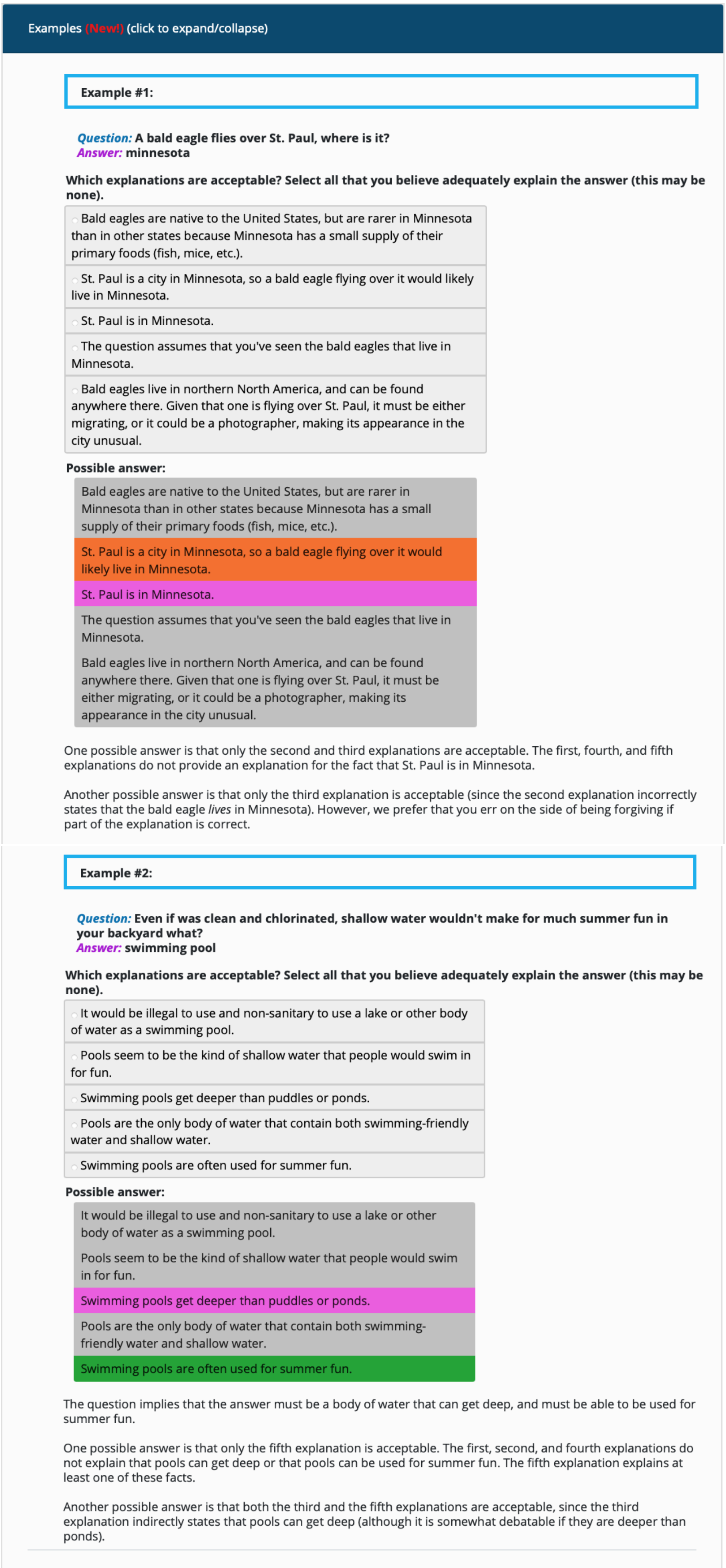


Figure 9: The examples given in the user interface of our explanation acceptability AMT studies for CommonsenseQA. The full interface is shown in Figure 8.