# Human Ratings Do Not Reflect Downstream Utility: A Study of Free-Text Explanations for Model Predictions

**Jenny Kunz**[1], **Martin Jirénius**[2], **Oskar Holmström**[1], **Marco Kuhlmann**[1]
Dept. of Computer and Information Science
Linköping University
[1]firstname.lastname@liu.se, [2]martin.jirenius@gmail.com

## Abstract

Models able to generate free-text rationales that explain their output have been proposed as an important step towards interpretable NLP for "reasoning" tasks such as natural language inference and commonsense question answering. However, the relative merits of different architectures and types of rationales are not well understood and hard to measure. In this paper, we contribute two insights to this line of research: First, we find that models trained on gold explanations learn to rely on these but, in the case of the more challenging question answering data set we use, fail when given generated explanations at test time. However, additional fine-tuning on generated explanations teaches the model to distinguish between reliable and unreliable information in explanations. Second, we compare explanations by a generation-only model to those generated by a self-rationalizing model and find that, while the former score higher in terms of validity, factual correctness, and similarity to gold explanations, they are not more useful for downstream classification. We observe that the self-rationalizing model is prone to hallucination, which is punished by most metrics but may add useful context for the classification step.

## 1 Introduction

Adding free-text explanations to NLP models is appealing as such explanations are easy to understand to human users and can include richer reasoning than methods that assign relevance scores to the input, such as LIME (Ribeiro et al., 2016) or saliency maps (Simonyan et al., 2014). Therefore, several commonsense reasoning data sets have been enriched with natural language explanations (Camburu et al., 2018; Rajani et al., 2019; Aggarwal et al., 2021). However, there is also significant scepticism, as the association between the model's predictions and its generated explanations is unclear. Bommasani et al. (2021) note that explanations may seem plausible but do not provide true insight into the model's reasoning, which fits the observation that open-ended generation models are prone to hallucinating unfaithful content (Maynez et al., 2020). Also, human explanations are not designed to be valid (or even complete) mechanisms leading to a correct prediction (Tan, 2022).

In this work, we study the effects of different design choices and properties of automatically generated explanations on the predictive performance of rationale-augmented models. To this end, we make targeted modifications to the model architecture and compare with gold-standard explanations. A common architecture for rationale-augmented models is a *pipeline* that maps the input to a rationale and the rationale to the output (I $\rightarrow$ R; R $\rightarrow$ O). Pipeline models are faithful by construction, but inferior in their performance. *Self-rationalizing models* that generate the rationale along with the output (I $\rightarrow$ OR) show good performance, but it is hard to assess the faithfulness of their explanations (Wiegreffe et al., 2021). We focus on a less-studied usage of free-text explanations, a rationale-enriched pipeline mapping the input to the rationale and the input along with the rationale to the output (I $\rightarrow$ R; IR $\rightarrow$ O). This architecture was originally proposed by Rajani et al. (2019) in their CAGE (Commonsense Auto-Generated Explanations) model. In the taxonomy of Hase et al. (2020), we are dealing with *serial-task reasoning models*. While not inherently faithful, as a causal path from input to predicted label remains open, these models allow us to study interactions between inputs and explanations more directly than self-rationalizing models because they allow for interventions at the explanation level, prior to the classification step. At the same time, Wiegreffe et al. (2021) show that the performance is superior to R $\rightarrow$ O, particularly when annotators are not instructed to provide self-contained explanations.

We use the framework of rationale-enriched pipelines to generate insights along two lines:



*Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164 - 177

1. We compare classification models solely trained on ground-truth explanations with models additionally fine-tuned on generated explanations. We find that the latter always perform notably better, while the former fail completely on the more challenging of our data sets.

2. We ask how explanations generated by a serial-task model (I $\rightarrow$ R) compare to those generated by a multi-task model (I $\rightarrow$ OR). We find that, while the serial-task explanations are more similar to gold explanations and their validity and factual correctness are ranked higher by human annotators, there is no clear difference in terms of utility for the classification step (IR $\rightarrow$ O).

## 2 Background

Annotating free-form explanations for NLP data sets has gained attention in recent years as language generation models became stronger. The popular natural language inference dataset SNLI (Bowman et al., 2015) has been enriched with crowd-sourced text explanations, resulting in e-SNLI Camburu et al. (2018). Two extensions were created for CommonsenseQA (Talmor et al., 2019), called CoS-E (Rajani et al., 2019) and ECQA (Aggarwal et al., 2021). In SemEval-2020 Task 4, a subtask was to generate a reason why a natural language statement does not makes sense to humans (Wang et al., 2020). Ling et al. (2017) solve algebraic word problems and generate a series of small steps necessary to derive the answer. Textual explanations have also been proposed for self-driving vehicles (Kim et al., 2018). The need for manual annotations of natural language explanations creates challenges, such as annotation costs (Belinkov and Glass, 2019). Also, human explanations can take various forms and have different goals (Miller, 2019) and do not necessarily verbalize valid reasoning paths (Tan, 2022).

### 2.1 Automatic Evaluation and Diagnostics

Two main characteristics are commonly included into the evaluation of explanations: Similarity with human-generated explanations and faithfulness towards the model's true decision-making process. Evaluating *extractive* explanations is straightforward at the first glance: If overlap with human importance assignments is desired, classical metrics such as $F_n$-scores can be used. Distinguishing between faithful and unfaithful explanations is harder, as there is no ground truth to compare to (Jacovi and Goldberg, 2020). Faithfulness is often evaluated by testing the model's performance after perturbing the input in relevant parts; see e.g. DeYoung et al. (2020) and Atanasova et al. (2020). The results obtained from such metrics are however not always consistent (Chan et al., 2022).

The evaluation of free-text explanations, which typically include input-external facts and reasoning, is a topic of ongoing discussion. Surface-level text generation metrics that measure the textual similarity of the generated explanation with the gold explanation have been employed, like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), which measure n-gram overlap, or BERTScore (Zhang et al., 2020), which sums cosine similarities between the BERT (Devlin et al., 2019) embeddings of the tokens in two sentences. BERTScore has been reported to correlate better with human judgement than other metrics in generation tasks (Zhang et al., 2020). The inconsistency of free-form explanations presents an obvious problem however, as there can be a large number of valid explanations that differ not only in surface form but also in reasoning paths. Also, humans and models may prefer different reasoning paths, resulting in a disconnect of generated explanations and model decisions.

To evaluate the faithfulness of explanations, Hase et al. (2020) suggest the Leakage-Adjusted Simulatability (LAS) metric, where the performance of a classifier with access to explanations is compared to its input-only version. In addition, they control for label leakage in the explanations by grouping data for which the label can be predicted solely with the explanation. Wiegreffe et al. (2021) show that a self-rationalizing T5 model (Raffel et al., 2020) fulfills two necessary conditions for faithful explanations: the robustness of output and explanations to input noise is correlated, and labels and rationales have a high feature importance agreement. While such approaches to evaluate the connection between explanations and predictions are insightful first steps, we are still scratching the surface. Evaluating faithfulness remains an unsolved problem.

### 2.2 Human Evaluation

While human evaluation is costly, it can provide important insights about properties such as factual correctness, which are not caught by automated metrics. A manual evaluation of explanation plausibility conducted by Marasovic et al. (2021) shows

that the qualitative difference of human and generated explanations remains substantial even with the largest available models. Wiegreffe et al. (2022) show that humans often prefer explanations generated by GPT-3 (Brown et al., 2020) over crowdsourced explanations. While the automatically generated explanations were rated low on qualitative criteria such as support of the label and novelty of information by default, a supervised acceptability filtering model based on human ratings of explanations improved explanation quality.

**Other Domains** Abstractive summarization is an insightful use case to evaluate generation faithfulness, as it is straightforward to judge if facts were in the original text. Maynez et al. (2020) show that the majority of summaries contain erroneous hallucinated content. Monsen and Rennes (2022) conduct a user study on abstractive versus extractive summaries. Their results show that abstractive summaries are much worse aligned with the meaning of the original text, resulting in factual incorrectness. Kryscinski et al. (2019) also report factual inconsistencies in a large number of abstractive summaries with a manual evaluation, and weak correlation between human ratings and ROUGE scores.

## 3 Experimental Setup

We generate and evaluate explanations in reasoning pipeline models using the following setups:

### 3.1 Data Sets

We use two English-language commonsense reasoning data sets that include human-annotated free-text explanations: ECQA and e-SNLI.

**ECQA** The Explanations for CommonsenseQA (ECQA) dataset (Aggarwal et al., 2021) extends the multiple-choice commonsense question answering data set CommonsenseQA (Talmor et al., 2019). For each question, five answer choices are provided. While Rajani et al. (2019) proposed the first extension of CommonsenseQA, their CoS-E data set has been reported to be of low quality: answers are ungrammatical (Narang et al., 2020) and rated exceptionally bad by humans (Wiegreffe et al., 2022). Explanations in ECQA are more detailed than in CoS-E. ECQA also includes refuting explanations for incorrect answer choices.

In our models, we provide one answer option with the respective explanation at a time, and use the target label *justify* if the answer is the correct one and *refute* if it is a wrong one. We create one training example for each annotated positive property and sample the data to get a ratio of 50/50 for positives/negatives during training.

**e-SNLI** The second data set we use is the natural language inference data set e-SNLI (Camburu et al., 2018). It is based on the popular SNLI (Bowman et al., 2015) that classifies the logical relation between a premise and a hypothesis sentence. It has three labels: *entailment*, *neutral* and *contradiction*. SNLI has been shown to contain annotation artifacts (label-specific lexical choices and the length of the hypothesis) that allow for correct classifications without solving the task (Gururangan et al., 2018), making explanation annotations to guide the model even more interesting. In fact, Camburu et al. (2018) show that correct explanations are much less likely to emerge from artifacts than correct labels. Explanations in e-SNLI are largely self-contained: Camburu et al. (2018) report that the classification accuracy conditioned only on the explanation is 96.83%.

### 3.2 Models

As previously mentioned, our reasoning models consist of a generator and a classifier. We implement all models on top of the PyTorch (Paszke et al., 2019) and Hugging Face Transformers (Wolf et al., 2020) libraries and follow standard fine-tuning strategies.[1]

#### 3.2.1 Generation Models

We use two models to collect explanations.

- Our single-task model (called **GPT-ST** in the following sections) is a GPT-2 (Radford et al., 2019) model that we fine-tune on the task-specific data using a language modelling head.
- The multi-task model (**GPT-MT**) is a GPT-2 model with *two* heads, one for language modeling and one for label classification. We use a weighted additive loss to combine the LM and the classification loss.

The prompt for the GPT-2 components is: "Statement: + *Question or Premise* + Statement: + *Answer Option or Hypothesis* + Explanation: + *Explanation*". We only account for tokens in the

[1]All code with dependencies and parameters is available at https://github.com/martinjirenius/reasoning-pipeline-models

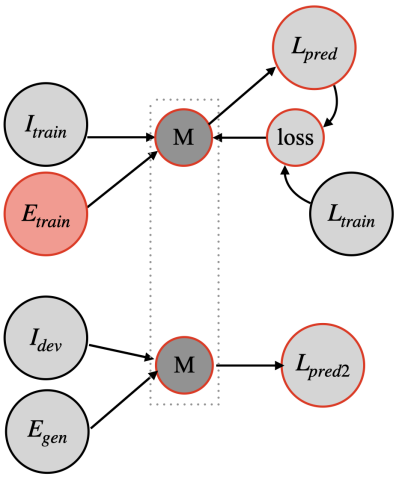


Figure 1: Experimental setup for training (upper half) and testing (lower half) on gold versus generated explanations as a causal graph (Pearl, 1995). $I_{train}$, $I_{dev}$, $E_{train}$, $L_{train}$ and $L_{dev}$ are the inputs, explanations and labels from the train and dev set, respectively. $E_{gen}$ are generated explanations from the GPT-models, $M$ is the BERT classification model, $L_{pred}$ are the labels predicted by $M$. All variables affected by the intervention on $E_{train}$ are marked with a red border line.

generated explanation when calculating the cross-entropy loss.

#### 3.2.2 Classification Models

For classification, we use fine-tuned BERT base models (Devlin et al., 2019) and present the input in the format "`[CLS]` + *Question or Premise* + `[SEP]` + *Answer Option or Hypothesis* + `[SEP]` + *Explanation* + `[SEP]`". We evaluate six different setups for each data set as specified in Table 1.

$\text{BERT}_{\text{none}}$ is a lower-bound baseline that does not use any explanations. $\text{BERT}_{\text{gold}}$ is an upper-bound baseline that uses gold explanations both for training and at test time. $\text{BERT}_{\text{ST}}$ uses gold explanations for training and the explanations generated by GPT-ST at test time. $\text{BERT}_{\text{ST-ft}}$ uses explanations of GPT-ST at test time, but different from $\text{BERT}_{\text{ST}}$ it is also fine-tuned on GPT-ST explanations. $\text{BERT}_{\text{MT}}$ uses gold explanations for training and the explanations from GPT-MT at test time. $\text{BERT}_{\text{MT-ft}}$ is fine-tuned and tested on GPT-MT explanations.

| | Trained with: | Tested with: |
|---|---|---|
| $\text{BERT}_{\text{none}}$ | – | – |
| $\text{BERT}_{\text{gold}}$ | Gold | Gold |
| $\text{BERT}_{\text{ST}}$ | Gold | GPT-ST |
| $\text{BERT}_{\text{ST-ft}}$ | GPT-ST | GPT-ST |
| $\text{BERT}_{\text{MT}}$ | Gold | GPT-MT |
| $\text{BERT}_{\text{MT-ft}}$ | GPT-MT | GPT-MT |

Table 1: Overview of our classification setups. The table indicates the source of the explanations that the model is trained and tested with.

Figure 1 illustrates the latter four models with regard to the intervention of fine-tuning the model on generated explanations, i.e. going from $\text{BERT}_{\text{ST}}$ to $\text{BERT}_{\text{ST-ft}}$ and from $\text{BERT}_{\text{MT}}$ to $\text{BERT}_{\text{MT-ft}}$.

### 3.3 Evaluation

**Quantitative Metrics** Our primary evaluation criteria are the similarity between the generated explanations and the gold explanations, as well as the predictive performance of the complete pipeline. To quantify the similarity, we use BERTScore (F1). To evaluate the classifiers, we compute their macro-averaged F1 score and accuracy on the test data.

Note that the native labels for our ECQA models are *justify* and *refute* for each possible answer. To make our evaluation comparable to other work, we calculate accuracy based on the answer with the highest score for *justify*.

**Human Evaluation** To assess qualitative properties of the generated explanations, we conduct a human evaluation over 200 random samples for each of the data sets. Inspired by the human evaluation studies by Monsen and Rennes (2022) and Wiegreffe et al. (2022), we ask annotators the following questions for each $e$ out of the gold, GPT-ST and GPT-MT explanations of each sample:

- Is $e$ a well-formed sentence?
- Does $e$ support the label?
- Is the content of $e$ factually correct?
- Does $e$ provide a valid reasoning path for the label?
- Does $e$ add new information, rather than recombining information from the input?

The possible answers for each question are *yes* or *no*. Each sample is rated by three persons familiar with the tasks (the first three authors). We report the average score across reviewers as well as Krippendorf's $\alpha$ ($n = 3$, interval from $-1$ to 1) for inter-rater agreement (Krippendorff, 2011). The full instructions for the annotators can be found in Appendix A. The data is available at `https://github.com/jekunz/bbnlp22_human`.

| | GPT-ST | GPT-MT |
|---|---|---|
| ECQA | **0.3108** | 0.2502 |
| e-SNLI | 0.3989 | 0.4009 |

Table 2: BERTScores (F1) for the single-task (GPT-ST) and multi-task (GPT-MT) models.

## 4 Results

We present our main results together with some additional follow-up experiments.

### 4.1 BERTScores and Surface Features

First, we test if GPT-ST or GPT-MT generates better explanations as evaluated by BERTScores. We see in Table 2 that GPT-ST explanations are more similar to the human reference explanations than GPT-MT solutions, at 0.3108 vs. 0.2502. For e-SNLI, the BERTScores for both models are very close, and much higher than those for ECQA, at 0.3989 resp. 0.4009.

We also compare the generated explanations in terms of simple surface features: explanation length, vocabulary size and vocabulary overlap with gold explanations (Table 3), and find that for e-SNLI, GPT-ST and GPT-MT explanations have almost identical characteristics. For ECQA, the difference is more substantial: While GPT-ST explanations are shorter than both GPT-MT and gold explanations, the former's vocabulary is larger than that of GPT-MT. The overlap with gold explanations is slightly higher for GPT-MT.

### 4.2 Classification

Results for the classification models are reported in Tables 4 and 5 (macro-averaged F1, accuracy).

| | GPT-ST | GPT-MT | Gold |
|---|---|---|---|
| ECQA: Words | 9.14 | 10.28 | 10.54 |
| ECQA: Chars | 48.08 | 49.74 | 57.48 |
| ECQA: Vocab | 7,946 | 4,436 | 11,033 |
| ECQA: Overl. | 0.772 | 0.735 | – |
| e-SNLI: Words | 11.79 | 11.77 | 13.32 |
| e-SNLI: Chars | 60.11 | 60.01 | 68.75 |
| e-SNLI: Vocab | 9,398 | 9,346 | 14,935 |
| e-SNLI: Overl. | 0.860 | 0.860 | – |

Table 3: Surface features: average word and character length, vocabulary size and vocabulary overlap with gold explanations for each set of explanations (dev. set).

**Baselines** As expected, the baseline $\text{BERT}_{\text{gold}}$ performs best across all metrics, models and data sets. For e-SNLI, $\text{BERT}_{\text{none}}$ performs better than all models that utilize generated explanations. For ECQA, $\text{BERT}_{\text{ST}}$ and $\text{BERT}_{\text{MT}}$ get a classification accuracy below the $\text{BERT}_{\text{none}}$ accuracy, with 0.253 and 0.231 compared to a random baseline of 0.2. However, looking at the F1 scores, we see that the $\text{BERT}_{\text{none}}$ baseline is outperformed by all ECQA explanation models.

**Fine-tuning on generated explanations improves results** When fine-tuning on generated explanations in the $\text{BERT}_{\text{ST-ft}}$ and $\text{BERT}_{\text{MT-ft}}$ models, the explanation models outperform the $\text{BERT}_{\text{none}}$ baseline for ECQA consistently, showing that the additional supervision with generated explanations is helpful. While for e-SNLI $\text{BERT}_{\text{none}}$ is not outperformed, the *ft* models still perform consistently better than the models trained on gold explanations, although the gap is smaller than for ECQA.

As an ablation, we also train two ECQA BERT models ($\text{BERT}_{\text{ST-abl}}$ and $\text{BERT}_{\text{MT-abl}}$) on generated explanations only, and evaluate them on gold explanations. $\text{BERT}_{\text{ST-abl}}$ achieves an accuracy of 0.522 on gold explanations and $\text{BERT}_{\text{MT-abl}}$ achieves 0.479, improving over comparable models that utilize generated explanations by at least 0.062 ($\text{BERT}_{\text{ST-ft}}$: 0.460) and 0.011 ($\text{BERT}_{\text{MT-ft}}$: 0.468). The accuracies of the ablation models on generated explanations are 0.406 ($\text{BERT}_{\text{ST-abl}}$) and 0.469 ($\text{BERT}_{\text{MT-abl}}$ ). Still, the gap to the gold-trained and gold-evaluated model remains substantial.

**Single-task versus multi-task explanations** While the BERTScore differences between GPT-ST and GPT-MT explanations are large for ECQA, using these explanations downstream in the classification model gives very similar results. For ECQA, the MT model even appears to have a slight advantage at least for the *ft* models, while for e-SNLI, it is the other way round.

### 4.3 Human Evaluation

The results of the human evaluation are reported in Table 6. In the case of ECQA, we see that the annotators have a preference for the GPT-ST explanations, giving them considerably higher scores for *support*, *correctness* and *validity*. The GPT-MT model adds more novel information. A closer look at the novel information shows that in the examples that were flagged to contain novel information, the majority (0.637) are factually incorrect. The

| | $\text{BERT}_{\text{none}}$ | $\text{BERT}_{\text{gold}}$ | $\text{BERT}_{\text{ST}}$ | $\text{BERT}_{\text{ST-ft}}$ | $\text{BERT}_{\text{MT}}$ | $\text{BERT}_{\text{MT-ft}}$ |
|---|---|---|---|---|---|---|
| ECQA | 0.378 | **0.906** | 0.514 | 0.631 | 0.489 | **0.634** |
| e-SNLI | 0.898 | **0.980** | 0.836 | **0.861** | 0.836 | 0.861 |

Table 4: Results for the classification models, macro-averaged F1 scores.

| | $\text{BERT}_{\text{none}}$ | $\text{BERT}_{\text{gold}}$ | $\text{BERT}_{\text{ST}}$ | $\text{BERT}_{\text{ST-ft}}$ | $\text{BERT}_{\text{MT}}$ | $\text{BERT}_{\text{MT-ft}}$ |
|---|---|---|---|---|---|---|
| ECQA | 0.338 | **0.945** | 0.253 | 0.460 | 0.231 | **0.468** |
| e-SNLI | 0.898 | **0.993** | 0.844 | **0.866** | 0.843 | 0.863 |

Table 5: Results for the classification models, accuracy.

annotators anecdotally report a large amount of nonsensical hallucinations in the GPT-MT model; we include examples in Appendix B.1. The overall scores are low, with shares of *yes* answers to the validity criterion being only 0.285 (GPT-ST) and 0.107 (GPT-MT). However, the gold answers do not get good scores either, with a *yes* share of 0.49. The highest-scoring criterion is *well-formedness*, where GPT-MT gets scores comparable to the gold explanations. With 0.607 vs. 0.603, the share of well-formed answers is however still low, with the generation models probably mirroring sloppy explanations in the training set.

For e-SNLI, the scores for all criteria except *novelty* are considerably higher. There is a slight preference for GPT-MT in the criteria *support*, *correctness* and *validity*, and a slight preference for GPT-ST in *well-formedness*, where GPT-ST even surpasses the gold explanations (0.868 vs. 0.833). Annotators noted that the ease of creating well-formed explanation may be due to the explanation often following clear templates; examples are given in Appendix B.2. e-SNLI explanations almost never add new information; the highest share is in the gold set with only 0.052.

For both data sets we note that the inter-annotator agreement on gold explanations is much lower than on both sets of generated explanations.

## 5 Discussion

We now discuss or results and method.

### 5.1 Results

**The downstream utility of explanations is not reflected by BERTScores or human ratings** The rationale-enriched pipeline helps us to better understand interactions between predictions and explanations by comparing the usefulness of different sets of explanations. Perhaps not surprisingly, we see that BERTScores do not reflect the usefulness of the explanations generated by different models. Large drops in BERTScores go along with at most very slight drops in the model's performance on the respective predictions. This is in line with results by Hase et al. (2020), who report that BLEU scores are not correlated with LAS.

Perhaps surprisingly however, the same effect is observed for the interplay of the human ratings and the downstream usefulness: Large differences in the human ratings of the validity and factual correctness of the explanations are not at all reflected in the downstream utility of the explanations. We hypothesize that a key property that leads to this behavior is the tendency of GPT-MT to hallucinate in ECQA (§ 4.3): While novel but factually incorrect information is punished in human ratings and BERTScore, the new information can still help the downstream model by adding possible context. GPT-ST on the other hand tends to "play safe" by creating more template-like explanations, with often sensible results but without novel information, and thereby without additional features for the classifier. Consider this example:

**Q:** The archaeologist was seeing artifacts that he knew were fake, how did he feel?
**A:** painful memories
**Label:** refute
**GPT-ST:** Painful memories is not a feeling.
**GPT-MT:** A person who is in fear of being embarrassed is called a bad person.

The GPT-ST explanation is reasonable but merely re-combines words from the question and answer. GPT-MT on the other hand creates an off-topic explanation that could, however, help the reasoning of the classifier by giving hints on alternative answers (like *embarrassed* or *fear*). We leave an investigation of this to future work.

| | Well-formed | Support | Correctness | Validity | Novelty |
|---|---|---|---|---|---|
| ECQA gold | 0.603 (+0.22) | 0.682 (+0.13) | 0.593 (−0.03) | 0.490 (+0.18) | 0.173 (+0.20) |
| ECQA GPT-ST | 0.573 (+0.25) | 0.513 (+0.45) | 0.443 (+0.19) | 0.285 (+0.48) | 0.126 (+0.28) |
| ECQA GPT-MT | 0.607 (+0.32) | 0.320 (+0.43) | 0.333 (+0.15) | 0.107 (+0.43) | 0.211 (+0.23) |
| e-SNLI gold | 0.833 (+0.04) | 0.873 (+0.06) | 0.860 (+0.08) | 0.772 (−0.06) | 0.052 (−0.02) |
| e-SNLI GPT-ST | 0.868 (+0.10) | 0.807 (+0.57) | 0.755 (+0.73) | 0.670 (+0.65) | 0.018 (+0.26) |
| e-SNLI GPT-MT | 0.830 (+0.24) | 0.813 (+0.56) | 0.813 (+0.56) | 0.688 (+0.54) | 0.012 (−0.01) |

Table 6: Human evaluation: average share of *yes* answers across all samples that were not flagged as invalid. The numbers in parentheses show Krippendorf's $\alpha$ ($n = 3$, interval from $-1$ to $+1$) for inter-rater agreement.

**Fine-tuning on generated explanations is crucial** Another important finding is the failure of $\text{BERT}_{\text{ST}}$ and $\text{BERT}_{\text{MT}}$ when encountering generated explanations in ECQA, which shows that our generator models do not catch the relevant semantic aspects sufficiently well for the classifier to rely on them. However, after fine-tuning with generated explanations, the BERT classifier can improve over the baseline without access to explanations. This shows that the model can still profit from the imperfect explanations if it learns to handle their limitations better. Our ablation with a model trained on generated and evaluated gold explanations suggests that it is not surface differences that make the transfer hard: The ablation model can in fact handle the gold explanations quite well, performing even better than on generated explanations. The fact that it still performs much worse than $\text{BERT}_{\text{gold}}$ on gold explanations shows that the model is far from perfect in identifying reliable information in the explanations; however, it is able to differentiate *to some extent*.

In previous work, Rajani et al. (2019) use a similar model consisting of GPT-2 and BERT, and succeed with gold-explanation training and generated-explanation testing for CoS-E. One reason for the contradictory results could be a more sophisticated optimization of their model, but we find it worth discussing that the success does not necessarily come by default. Another hypothesis is that the cause is the (reportedly) low-quality annotations in CoS-E (Narang et al., 2020) having a similar noise-adding effect as the generated explanations, and therefore allow the model to transfer.

**e-SNLI is easy, ECQA problematic to explain** On e-SNLI, all models get higher scores in all metrics than on ECQA. The only exception is novelty in the human evaluation: Novel information is not necessary to explain e-SNLI instances; it is sufficient to re-combine parts of premise and hypothesis. This is commonly done in a template-like manner:

– *[Part of premise] is [part of hypothesis]* for the *entailment* label,

– *Not all [part of premise] are [part of hypothesis]* for *neutral*, and

– *[Subject] cannot [part of premise] and [part of hypothesis] at the same time* for *contradiction*.

For full examples containing these patterns, we refer to Appendix B.2. The template-like explanations in e-SNLI have also been noted by Camburu et al. (2018) and Brahman et al. (2021). Such observations could raise the question if templates could be a more appropriate form of explanation for this data set, as they would improve clarity and reliability. Wiegreffe and Marasovic (2021) review explanation data sets and question the popular perception that template-like explanations are generally dismissed as uninformative. The authors suggest to instead embrace naturally occurring structures.

ECQA explanations rarely follow simple patterns and more often include external information. The low *validity* scores even for the gold explanations show that the data set is rather hard to explain. Our annotators noted that "incorrect" answer options in ECQA are not generally implausible but often just less likely than the "correct" option. This makes it hard to write explanations that do not explicitly consider the correct answer option in a contrastive manner (arguing why it is more likely than the current candidate). Examples are given in Appendix B.3. ECQA contains a notable number of uninformative explanations for the *refute* label both in the gold and the generated explanations, e.g. *[Answer] is not a correct option* (see Appendix B.4 for examples). This is possibly a result of annotators not being able to formulate satisfying reasons why the answer option is incorrect. ECQA also has a large amount of ungrammatical and low-quality annotations, which affects the generation models negatively.

### 5.2 Limitations

We conclude this section with a discussion of the limitations of our study.

**Model** An obvious limitation of our work is that our results on SNLI and CommonsenseQA are below the current state of the art, due to the moderate size of our models. While combining GPT-2 and BERT is a common setup for free-form explanation generating models (Wang et al., 2020), Hase et al. (2020) report much higher results using T5 models, and Marasovic et al. (2021) clearly document the effect of scale in a few-shot setup, with e-SNLI climbing from 79.2% to 87.4% and ECQA from 41.4% to 65.9% in classification accuracy when going from T5-base to T5-3B. While repeating the experiments with larger models could lead to different conclusions, we believe that investigating the smaller, more accessible and widely used models remains valuable.

**Evaluation** Another limitation in our analysis is the possibility that the multi-task explanations are affected by error propagation when the system makes wrong predictions.[2] This issue may affect both BERTScores and human evaluations. We suggest that a promising fix to this potential problem is to over-generate explanations and randomly choose one that accompanies a correct prediction.

**Data sets** That explanations do not increase the overall performance of SNLI models is known in the literature. Camburu et al. (2018) report a decline in accuracy with explanations: 84.01% for SNLI, but 83.96% for the best explanation model. Note that their models were BiLSTM models trained from scratch, as their work preceded current pre-trained models. Another work reports an improvement in accuracy, but with 0.3% it is extremely slight (Zhao and Vydiswaran, 2021). As pre-trained models get a superhuman performance on SNLI, and because of the known presence of annotation artifacts (Gururangan et al., 2018), recent improvements may however not be meaningful for solving the actual task. In addition, the high performance of models is not aligned with human agreement on natural language understanding tasks. In a human evaluation of SNLI by (Bowman et al., 2015), all annotators agree only on 58% of the labels.

[2]This limitation was rightfully noted by one of the reviewers of this paper, which we gratefully acknowledge.

Both data sets we use consist of crowd-sourced explanations of mixed quality. Doing a manual inspection of either of them, it is easy to find incorrect and logically inconsistent explanations, or explanations that contribute no additional information (§§ B.3, B.4). Our low inter-annotator agreement on gold explanations is an indicator of these problems. Related observations have also been raised in previous evaluations (Wiegreffe et al., 2022). Besides data quality, the tasks of natural language inference and multiple-choice question are arguably artificial. It is unclear how the results would transfer to explanation generation in general.

**The status of free-text explanations** We believe it is appropriate to remain sceptical about the utility of generated free-text explanations. Large models produce better explanations by all metrics, but there is still a huge qualitative difference of human and generated explanations (Marasovic et al., 2021). The acceptability filtering system proposed by Wiegreffe et al. (2022) improves human ratings of model-generated explanations substantially, but may, as these authors state themselves, be more relevant for goals such as creating trust in the system than for creating explanations faithful to the model's prediction process. In fact, generating explanations without guarantees of a causal connection between explanation and label is not faithful, and evidence that there is such a connection is sparse. Still, while we would strongly advise against using generated explanations as evidence about how a prediction was made, we argue that they can generate valuable insights into the "reasoning" capabilities of models, and thereby help improving models, task formulations and data sets. Unfortunately, the current lack of high-quality annotated data sets with explanations for diverse tasks makes it hard to fully assess their potential.

## 6 Conclusion

In this paper we compared free-text explanations in variants of a rationale-enriched pipeline: using a single-task versus a self-rationalizing generation model, and training the classifier on gold explanation only versus doing further fine-tuning with generated explanations. An extensive evaluation with similarity-based metrics, utility in downstream classification, and human ratings based on five different criteria shows limitations but also chances of free-text explanations. We see indications that hallucinations occur more frequently in explanations

by a self-rationalizing generation model. However, they do not appear to be generally harmful, and may even be useful for downstream predictions in rationale-enriched pipelines if the classification model has the chance to learn to differentiate between reliable and unreliable information. Further investigation of hallucinations in rationale-enriched pipelines, e.g. with extractive explanation methods, is an interesting avenue for future research.

That human ratings do not reflect classification utility indicates that it is crucial to design annotations and models targeted towards a use case: Explanations that convince human raters are not ideal for the goal of performance improvements by providing useful guidance to the model. However, the latter goal is not explicitly accounted for in popular data sets, but the former is not sufficiently met either, as particularly for ECQA, human annotators rate gold explanations low. Specialized explanations that maximize one goal at a time would help us understand the differences between human and model "reasoning", and thereby allow us to move towards more faithful free-text explanations.

## Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We thank the anonymous reviewers for their valuable and constructive feedback that contributed to improving this work.

## 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.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12592–12601.

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*, volume 31. Curran Associates, Inc.

Chun Sik Chan, Huanqi Kong, and Liang Guanqing. 2022. A comparative study of faithfulness metrics for model interpretability methods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5029–5038, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

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.

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578.

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167, Vancouver, Canada. Association for Computational Linguistics.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew E. Peters. 2021. Few-shot self-rationalization with natural language prompts. *CoRR*, abs/2111.08284.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Julius Monsen and Evelina Rennes. 2022. Perceived text quality and readability in extractive and abstractive summaries. In *Proceedings of the 13th international conference on Language Resources and Evaluation (LREC), Marseille, France*.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *CoRR*, abs/2004.14546.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

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. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188, Seattle, United States. Association for Computational Linguistics.

Cunxiang Wang, Shuailong Liang, Yili Jin, Yilong Wang, Xiaodan Zhu, and Yue Zhang. 2020. SemEval-2020 task 4: Commonsense validation and explanation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 307–321, Barcelona (online). International Committee for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable natural language processing. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xinyan Zhao and VG Vinod Vydiswaran. 2021. Lirex: Augmenting language inference with relevant explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14532–14539.

## A Full Instructions to Human Annotators

In the following box we provide the full instructions given to the human annotators for the manual evaluation. As all of our annotators are Master students or doctoral students who work with NLP, they were familiar with the e-SNLI and the ECQA data sets, so that we only needed to clarify the criteria and provide specific instructions.

If you disagree with the label or find the example to be non-sense: Flag the example with *N/A*.

**Step 1:** Look only at e internally:

- Is $e$ a well-formed sentence? ($e$ is grammatical and structurally sensible.)
- Is the content of $e$ factually correct? ($e$ itself is a true statement about the real world. $e$ is factually and logically correct.)

**Step 2:** Look at $e$ and the label:

- Does $e$ support the label? (Looking at e alone, it is reasonable that the label is correct.)

**Step 3:** Use all available context:

- Does $e$ provide a valid reasoning path for the label? ($e$ convincingly explains how to get from the context to the label.)
- Does $e$ add new information? (Rather than re-combining information from the context, $e$ comes up with new information.)

# B Anecdotical Examples for Observed Characteristics

In this appendix section we provide further examples for four phenomena mentioned in this paper that raters noticed during the human evaluation. These phenomena were not systematically reviewed at evaluation time but are anecdotal observations.

## B.1 Hallucinations in GPT-MT

The following five examples from GPT-MT include novel, "hallucinated" information that has no connection to the question or answer option.

**Q:** When old conservative people observe a changing society, what do the experience?
**A:** acceptance
**Label:** refute
**GPT-MT:** The experience of dying is a feeling

**Q:** Where can I put a coin to stop my car from being towed?
**A:** piggy bank
**Label:** refute
**GPT-MT:** A person can put a coin in a cowgy bank to stop his car from being washed from the water.

**Q:** After I urinate and flush the toilet and wash my hands, what should I do next?
**A:** zip pants
**Label:** refute
**GPT-MT:** A person should not put bag under shirt.

**Q:** The stud bull was brought in to breed with the prize winning what?
**A:** female
**Label:** refute
**GPT-MT:** Cats are not called horses

**Q:** Who might wear dark glasses indoors?
**A:** ray charles
**Label:** refute
**GPT-MT:** A person might wear dark glasses in a ring in the house.

## B.2 Template-like explanations in e-SNLI

In the following we provide six examples from e-SNLI, two for each label, where the explanations follow a clear template that is consistently followed by all explanation models.

**Premise:** A dog biting a man's rear
**Hypothesis:** An animal is biting a man.
**Label:** entailment
**GPT-ST:** A dog is an animal.
**GPT-MT:** A dog is an animal.
**Gold:** A dog is an animal.

**Premise:** A group of men playing music, the man holding a guitar is talking to a man with drums.
**Hypothesis:** A band is playing music.
**Label:** entailment
**GPT-ST:** A group of men playing music is a band.
**GPT-MT:** A group of men playing music is a band.
**Gold:** A group playing music together is know as a band.

**Premise:** Man holding torch of fire.
**Hypothesis:** The fire is large.
**Label:** neutral
**GPT-ST:** Not all fires are large.
**GPT-MT:** The fire is not necessarily large.
**Gold:** Not all the fires are large.

**Premise:** A woman in a bright pink and green skirt carries a large wicker basket on her back up a stone-cobbled street.
**Hypothesis:** The woman is carrying a basket down the street in Nepal.
**Label:** neutral
**GPT-ST:** Not all stone-cobbled streets are in Nepal.
**GPT-MT:** Not all stone-cobbled streets are in Nepal.
**Gold:** Just because a woman is carrying a basket does not mean that she is in Nepal.

**Premise:** A woman in a teal apron prepares a meal at a restaurant.
**Hypothesis:** A woman is walking in park
**Label:** contradiction
**GPT-ST:** The woman cannot be walking and preparing a meal at the same time.
**GPT-MT:** The woman cannot prepare a meal at a restaurant and walk in a park at the same time.
**Gold:** A restaurant is not a park.

**Premise:** A man in a blue shirt, blue jeans, and a red apron standing at a grill.
**Hypothesis:** A man bakes a cake.
**Label:** contradiction
**GPT-ST:** A man cannot bake a cake and stand at a grill at the same time.
**GPT-MT:** The man cannot be standing at a grill and baking a cake at the same time.
**Gold:** Standing at a grill cannot be done simultaneously with baking a cake.

### B.3 Plausible but "incorrect" answer options

The answer options in the following five examples are plausible, but not the most likely option. Reasons include too general or rather far-fetched but still possible options.

**Q:** Sean was a native English speaker, a fact that he found helpful. Ludwig didn't speak the language well because he was what?
**A:** alien
**Label:** refute
**GPT-ST:** Alien is not a person.
**GPT-MT:** A person cannot speak the language well if he is an island.
**Gold:** A human being is not an alien.

**Q:** Billy found that his toilet was overflowing. Water was splashing onto the floor and making a huge mess. He got a mop and tried to clean it up. Where might the toilet be located?
**A:** apartment
**Label:** refute
**GPT-ST:** Apartment is a community having 10 to 20 homes or a block of apartments. Apartment is not where toilet overflowed.
**GPT-MT:** Apartment is a room or set of rooms or a building where people live permanently, especially as a member of a family or household.
**Gold:** His toilet was not in the apartment.

**Q:** They really needed the information prisoner, so they did the least human thing and began to what the prisoner?
**A:** hurt
**Label:** refute
**GPT-ST:** Prisoners do not hurt when they are being hurt
**GPT-MT:** When a person is hurt, he gets hurt.
**Gold:** Hurt is not the least human thing they can do.

**Q:** When old conservative people observe a changing society, what do the experience?
**A:** acceptance
**Label:** refute
**GPT-ST:** Acceptance is not related to changing society
**GPT-MT:** The experience of dying is a feeling
**Gold:** Acceptance is not an emotion

**Q:** How can one's views change after learning about science?
**A:** appreciation of nature
**Label:** refute
**GPT-ST:** One can't change his views after learning about science.
**GPT-MT:** A person can change his ideas by learning about science.
**Gold:** Appreciation of nature is the act of recognising or understanding that nature is valuable, important or as described. Appreciation of nature is the act of recognising that nature is valuable while learning about science is all about greater knowledge and not appreciation.

### B.4 Uninformative "refute" answers

These five examples show explanations that do not contain incorrect facts, but are completely uninformative and therefore unfit as explanations.

**Q:** What can happen as a result of a business spending money and having to make a declaration?
**A:** poverty
**Label:** refute
**Gold:** Poverty is not an appropriate option.

**Q:** What animal has quills all over it?
**A:** calligraphy
**Label:** refute
**Gold:** Calligraphy dont suits in this sentence.

**Q:** What is unlikely to get bugs on its windshield due to bugs' inability to reach it when it is moving?
**A:** car
**Label:** refute
**GPT-MT:** Car is not a correct option.

**Q:** There was a cloud or two out but overall it was very what?
**A:** bringing rain
**Label:** refute
**GPT-MT:** Rain is not a correct option.

**Q:** Sam tried to go to sleep, but he couldn't. He suffered from what?
**A:** bed
**Label:** refute
**GPT-ST:** Bed is not a correct option.

**Q:** The coach decided to make a lineup change, the team's effort was suffering from what?
**A:** paper money
**Label:** refute
**GPT-ST:** Paper money is not a correct option.