What if you said that differently?: How Explanation Formats Affect Human Feedback Efficacy and User Perception

Chaitanya Malaviya, Subin Lee, Dan Roth, Mark Yatskar
University of Pennsylvania
{cmalaviy,subinlee,danroth,myatskar}@upenn.edu

Abstract

Eliciting feedback from end users of NLP models can be beneficial for improving models. However, *how should we present model responses to users so they are most amenable to be corrected from user feedback*? Further, what properties do users value to understand and trust responses? We answer these questions by analyzing the effect of rationales (or explanations) generated by QA models to support their answers.

We specifically consider decomposed QA models that first extract an intermediate rationale based on a context and a question and then use solely this rationale to answer the question. A rationale outlines the approach followed by the model to answer the question. Our work considers various formats of these rationales that vary according to well-defined properties of interest. We sample rationales from language models using few-shot prompting for two datasets, and then perform two user studies. First, we present users with incorrect answers and corresponding rationales in various formats and ask them to provide natural language feedback to revise the rationale. We then measure the effectiveness of this feedback in patching these rationales through in-context learning. The second study evaluates how well different rationale formats enable users to understand and trust model answers, when they are correct. We find that rationale formats significantly affect how easy it is (1) for users to give feedback for rationales, and (2) for models to subsequently execute this feedback. In addition, formats with attributions to the context and in-depth reasoning significantly enhance user-reported understanding and trust of model outputs.[1]

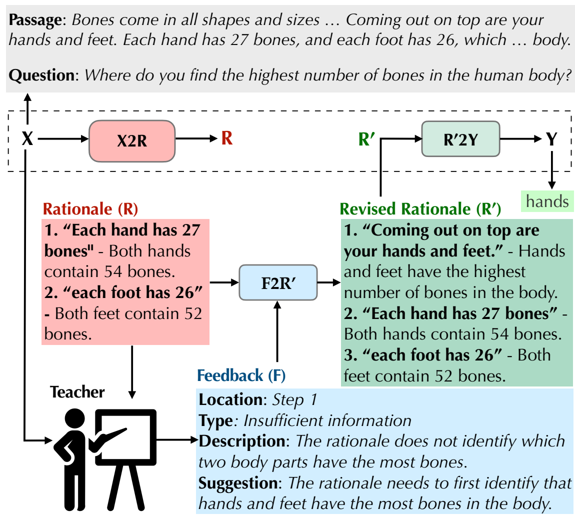


Figure 1: The framework for incorporating human feedback into decomposed QA models. A model $X2R$ generates a rationale R to answer a question based on a passage. A human teacher then provides natural language feedback for R, which is used to generate a revised rationale R' from $F2R'$. Finally, this revised rationale is used to generate the final answer Y. We study various formats of the intermediate rationale R.

1 Introduction

Question answering models can often be black boxes, as their reasoning process is mostly opaque to model builders as well as end users. This can inhibit the ability of users to provide helpful critiques to models to repair them. Generating *rationales* (or explanations) along with answers is a viable approach that can alleviate these concerns, but these rationales are inherently not faithful and can sometimes be inconsistent with the answers themselves (Ye and Durrett, 2022; Turpin et al., 2023; Lanham et al., 2023; Radhakrishnan et al., 2023).

This motivates approaches that decompose the question answering task into two stages (depicted with dashed lines in Figure 1), where we first generate a rationale for the question using the given context ($X2R$), then use only this rationale to answer the question ($R2Y$) (Lei et al., 2016; Eisenstein et al., 2022). A rationale may provide a justification for the answer by presenting an outline for how the question can be answered. By only relying

[1]Code and data is available at https://github.com/chaitanyamalaviya/rationale_formats.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3046–3065
June 16-21, 2024 ©2024 Association for Computational Linguistics

on the rationale as context, the answer generation model ($R2Y$) has a stronger inductive bias to generate an answer that is faithful to the rationale.

Crucially, faithful rationales can allow users to follow the model's line of reasoning, and subsequently provide actionable feedback to the model. This feedback can be used to repair individual outputs or enable generalization to novel instances. However, it is unclear precisely how a rationale should be formatted, i) to best aid the user's understanding of the model's reasoning, and ii) their ability to provide feedback for the response.

Our work specifically addresses the question of how we can format intermediate rationales (R) for decomposed QA systems, such that they are *easy to repair through human feedback*. Further, we analyze what properties make rationales *interpretable*, and *trustworthy* to users. Previous work on decomposed question answering mostly consider rationales as text snippets extracted from the context, optionally marked up with coreferences that make the snippets standalone (DeYoung et al., 2020; Eisenstein et al., 2022). Although extractive snippets can serve useful for providing minimal context that rationalizes an answer, they do not provide much insight into the model's reasoning process. This may limit a user's understanding and their ability to critique the model. We consider alternative formats of rationales, which vary according to well-defined characteristics (§3). Examples of these formats and how they vary are presented in Table 1.

Based on the considered rationale formats, we generate responses (rationales & answers) from a decomposed QA system. We then perform two user studies where we measure i) the effectiveness of user feedback in patching rationales in different formats and ii) the ability of different formats to enable users to understand and trust responses. In our first study, we sample *incorrect* answers corresponding to all rationale formats, and ask annotators to provide natural language feedback for the rationales (§5). We use this feedback to then revise the rationales ($F2R'$) and regenerate the final answer ($R'2Y$). Comparing the regeneration accuracy with different rationale formats provides insight into properties of rationales that make them easy to repair. Further, they provide an upper bound for how much improvement can be expected through automated feedback by repairing rationales (Chen et al., 2023; Madaan et al., 2023). In our second study, we elicit judgements of interpretability and trustworthiness for *correct* answers and their accompanying rationales from users (§6).

We consider two tasks: general reading comprehension (Quoref; Dasigi et al. (2019)) and medical reading comprehension (PubMedQA; Jin et al. (2019)). Each block in our overall workflow (Figure 1) is implemented through few-shot prompting of a large language model. Our experiments suggest that rationale formats significantly affect i) users' ease of providing feedback and ii) the model's ability to execute that feedback. In addition to being critiquable, certain rationale formats are more helpful in aiding users to understand and trust the model's answers. One such format is the *annotated report*, which includes a list of extractive phrases and free-text inferences based on each phrase. Finally, among a few characteristics of rationales presented to users, we find that users rate attribution and depth of reasoning as characteristics that are most important to them.

2 Problem Formulation

Consider a standard reading comprehension task, where we are given a passage, a question based on this passage, and a reference answer Y as a labeled example in our dataset. We represent the input information in the task (the passage and the question) with X. Typically, we would train a model $X2Y$ that predicts Y given X by learning $P(Y|X)$. Assuming $X2Y$ is a black-box model, without loss of generality, the model may internally compute a latent representation, which is usually not extractable in an interpretable format, from X to predict Y, internally decomposing the problem. This can restrict the transparency of the model because we cannot provide a faithful reasoning to an end user that supports the answer.

2.1 Decomposed QA Pipeline

In the decomposed QA pipeline, we factor the QA problem in the following manner (see Figure 1 for an illustration):

$$P(Y, R|X) = P(Y|R, X)P(R|X) \qquad (1)$$

Since Y is independent of X given R (assume Q is part of R for simplicity), we have,

$$P(Y, R|X) = P(Y|R)P(R|X) \qquad (2)$$

Let's first consider a rationale-generating model $X2R$, that generates a textual rationale R given X.

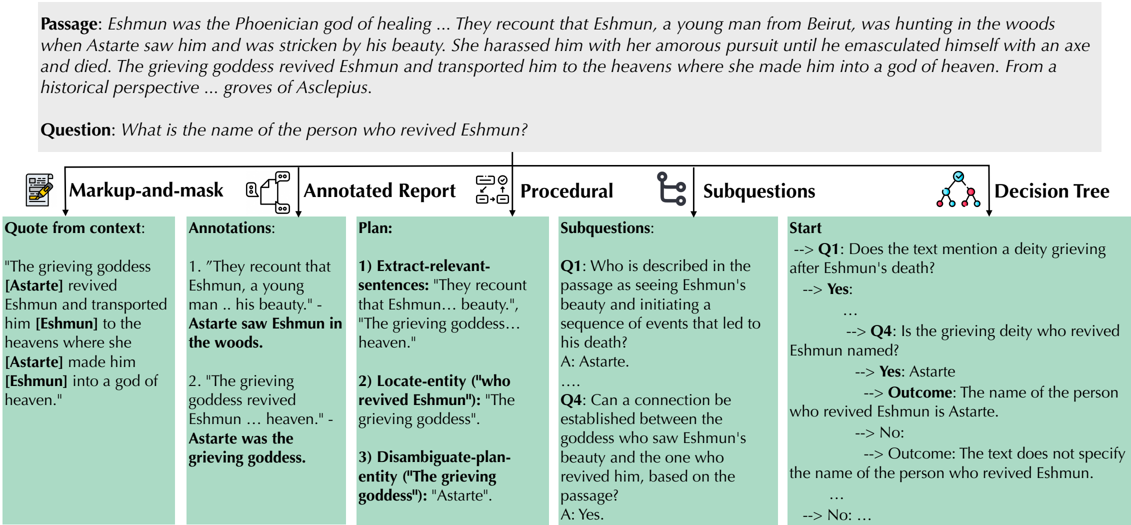


Figure 2: Examples of the different rationale formats considered for representing intermediate rationales.

This rationale provides an outline of the approach proposed by the model to answer the question. Consider also an answer-generating model $R2Y$ that generates an answer $\hat{Y}$ given the predicted rationale R. In this decomposed model, $R2Y$ has a strong inductive bias to use the information presented in R for its reasoning. Further, R can be explicitly shown to an end user, which increases the transparency of the entire system.

2.2 Debugging the Decomposed QA Pipeline

Next, let's assume a set of spans X_s from X that are sufficient to predict the answer Y. First, we note that for the answer to be correct, i.e. $\hat{Y} = Y$, the predicted rationale R must contain all the information contained in X_s, i.e., $I(X_s)$. Errors in answers generated by $R2Y$ can be a result of (1) insufficient or incorrect context, when $I(X_s) \not\subseteq I(R)$, and / or (2) limited model capacity of $R2Y$, when $I(X_s) \subseteq I(R)$. Repairing the modeling pipeline (i.e., $X2R + R2Y$) can hence either involve improving the quality of generated rationales R produced by $X2R$ or improving the modeling capacity of $R2Y$. Note that the example may require information beyond what is in the passage (for example, domain knowledge or commonsense knowledge).

We consider the scenario where we repair the rationales generated by $X2R$ through natural language feedback. We assume a teacher T who writes feedback for generated rationales R, where they describe flaws in R. Generated rationales can be lacking in various ways: 1) insufficient information: R may not contain crucial information required to perform the inference (i.e., $I(X_s) \not\subseteq I(R)$), or 2) incorrect information: R may contain hallucinated content or incorrect reasoning chains that could mislead the answer-generating model.

The teacher T in our case is an end user, who could optionally be a domain expert depending on the task. We evaluate whether the format of predicted rationales R is interpretable and easy to repair for T. Based on T's feedback F_k for a subset of examples $X_k \subset X$, we revise the initial rationale R to generate a revised rationale R' using a model $F2R'$. This revised rationale is then used by $R'2Y$ to generate the final answer.

3 Intermediate Rationale Formats

Rationales in NLP tasks are usually presented as compressed text snippets extracted from the given input (DeYoung et al., 2020; Eisenstein et al., 2022). However, text snippets from the context alone may not make the model's reasoning explicit and transparent to users. We consider alternative rationale formats that describe the model's reasoning. We describe these formats below. A summary of these formats and how they vary according to rationale properties, is given in Table 1.

3.1 Rationale Formats

Markup-and-Mask (`markup_mask`). This format, proposed by Eisenstein et al. (2022), extracts sentences from the context that are relevant to answering the question. Sentences are decontextualized by markups that resolve coreferences and other ambiguous phrases (Choi et al., 2021).

Format	Description	Attribution Provided	Reasoning Exposed	Sequential Reasoning	Free-text annotations
Markup-and-Mask	Quoted sentences from the context, marked up with coreferences for pronouns and ambiguous phrases	✓	✗	✗	✗
Annotated Report	Quoted phrases from the context and an inference from each phrase	✓	✓	✓	✓
Procedural	Step-by-step plan for solving the question with pre-defined operations	✓	✓	✓	✗
Subquestions	Breakdown of the original question into subquestions	✗	✓	✓	✓
Decision Tree	Breakdown of the original question into subquestions presented in a tree structure, with Yes/No outcomes for each subquestion	✗	✓	✓	✓

Table 1: Descriptions of the rationale formats considered in our work and the characteristics along which they differ.

Annotated Report (`annotated_report`). The annotated report extracts phrases from the context and generates a free-text inference based on each phrase that is relevant to answering the question. This is broadly inspired by the way readers highlight and annotate key spans in documents while reading (also found as marginalia in books).

Procedural (`procedural`). A procedural rationale is a step-by-step plan that uses predefined operations to answer the question. Similar works that broadly propose a plan-based rationale have been explored in prior work, in different contexts (Sun et al., 2023; Wang et al., 2023). The primitive operations we consider include an operation to extract relevant sentences, disambiguate an entity from the question or the plan so far, and locate an entity by answering a subquestion. These are further described in Appendix A.

Subquestions (`subquestions`). Subquestions simply decompose the original question into multiple questions that provide relevant information to answer the question. These have been explored as a form of rationale in various works (Geva et al., 2021; Khot et al., 2021; Press et al., 2022; Dua et al., 2022; Zhou et al., 2023a).

Decision Tree (`decision_tree`). We also consider a tree-structured rationale, inspired by fast-and-frugal trees (Martignon et al., 2003) as well as prompting work that explores tree-like structures (Yao et al., 2023). This format decomposes the original question into Yes/No subquestions in a tree-like structure and also shows the incorrect tree traversals for completeness.

4 Experimental Setup

4.1 Datasets

We consider two datasets for our studies: Quoref (Dasigi et al., 2019) and PubMedQA (Jin et al., 2019). The first is a general reading comprehension dataset while the second involves medical reading comprehension. In contrast to Quoref, PubMedQA often requires domain-specific knowledge for answering the question. We use all validation set examples of Quoref (2418 examples) and all labeled examples in PubMedQA (1000 examples)

4.2 Sampling Rationales and Answers

We sample rationales and answers for all 5 formats in a decomposed QA pipeline, where both $X2R$ and $R2Y$ are implemented using few-shot prompting. We first prompt `gpt-3.5-turbo` for rationales by providing the passage and question. The question and only the generated rationales are then used to prompt the same model to generate the final answer. In both cases, we sample few-shot exemplars using BM25 (Robertson et al., 2009) from a set of 100 manually labeled examples with rationales. We sample as many exemplars as can fit within the maximum sequence length (4096) of the model. This usually amounts to 3-5 exemplars. The prompts used and other hyperparameters are provided in Appendix A.

5 Study 1: Repairing Rationales through Human Feedback

5.1 Setup

In this study, we measure the critiquability or ease of repair of different rationale formats. This is done by collecting natural language feedback from human annotators for rationales corresponding to incorrect answers. We sample examples for which the decomposed QA pipeline predicts *incorrect answers* for all 5 rationale formats. In all, we collect 490 feedback statements for Quoref and 555 feedback statements for PubMedQA.[2]

In each annotation task, annotators are shown a single example (question & passage) with all 5 rationale formats and their corresponding answers. This controls for annotator variance and variance across examples. To control for ordering effects, we randomize the order in which the rationale formats are presented to annotators. For each rationale format, annotators are asked to write natural language feedback to repair the rationale.

We use this natural language feedback to generate the revised rationale R'. To do this, we prompt `gpt-3.5-turbo` with the passage, the question, the original rationale, and human-written feedback. Finally, we generate the final answer by few-shot prompting the same model using just the revised rationale and question. Prompts and other hyperparameters are in Appendix A.

5.2 Participants

For this study as well as the study in section 6, we recruit participants through Prolific. Participants are required to be fluent in English and are based primarily in English-speaking countries. For PubMedQA examples, they are required to be working in the healthcare sector. For more details, please see Appendix B.

5.3 Task

To prime annotators for formulating their feedback, we ask them to first evaluate the sufficiency and faithfulness of each rationale to the context. They label these two properties for each rationale format on a Likert scale (the precise descriptions of the options in all Likert-scale questions are in Figure 6).

Sufficiency. Annotators evaluate if the rationale provides enough information to answer the question, without the context. Note that the rationale may contain inaccuracies but still be sufficient. They mark sufficiency as (*Sufficient*, *A bit insufficient*, *Entirely insufficient*).

Faithfulness to context. Next, annotators evaluate whether the rationale accurately draws conclusions from the context without misrepresenting any information. They mark faithfulness on a scale of (*Accurate*, *A bit inaccurate*, *Very inaccurate*).

[2] This corresponds to 5 rationale formats and 98 examples for Quoref and 111 examples for PubMedQA.

	Quoref		PubMedQA
Rationale Format	**EM**	**F1**	**Accuracy**
`none`	70.31	79.65	69.30
`markup_mask`	57.44	68.10	62.20
`annotated_report`	60.26	70.20	70.20
`procedural`	66.09	77.05	68.30
`subquestions`	54.26	63.05	68.90
`decision_tree`	68.61	77.09	46.70

Table 2: Initial scores using the decomposed QA pipeline ($X2R$ + $R2Y$) for different rationale formats.

5.3.1 Feedback

Annotators are asked to formulate natural language feedback that would be most useful in directing the model to the reference correct answer. It could target missing or incorrect information in the rationale, but cannot explicitly reveal the correct answer. Feedback is elicited in multiple steps (examples of feedback written by annotators are in Table 4, 5):

1. **Location of error**: Annotators are required to list the step(s) (or question number) in which the error occurs.
2. **Type of error**: Annotators then identify the type of the error. We show them a few common error types that occur in rationales (for example, insufficient information, irrelevant information, incorrect inferences etc).
3. **Description of error**: Next, annotators use concrete details from the rationale, question & context to provide a description of the error.
4. **Actionable suggestion**: Finally, annotators provide an actionable edit that would repair the rationale, again using concrete details from the rationale, question & context.

Ease of repair. Using annotator feedback, we can measure how amenable each rationale format is for repair. However, this does not reflect annotators' ease of providing feedback for each format. We elicit this directly on a scale of (*Very easy*, *Somewhat easy*, *Somewhat hard*, and *Very hard*).

5.4 Evaluation

We first evaluate the effectiveness of feedback through edit accuracy (`edit_acc`), where we manually label each revised rationale for whether it incorporates the feedback. Next, we compute final answer accuracy (`final_acc`), where we check

	Quoref			PubMedQA		
Rationale Format	**edit_acc**	**final_acc**	**time_taken(s)**	**edit_acc**	**final_acc**	**time_taken(s)**
markup_mask	50.00	$29.69^{*\|*\|*\|*}$	397.38	71.03	$14.95^{*\|*\|o\|o}$	379.07
annotated_report	51.67	$38.33^{*\|o\|*\|o}$	381.52	62.96	$\mathbf{20.37}^{*\|*\|*\|*}$	426.04
procedural	57.89	$\mathbf{38.60}^{*\|o\|o\|*}$	359.75	55.77	$8.65^{*\|*\|*\|*}$	434.24
subquestions	49.21	$36.51^{*\|*\|*\|o}$	375.45	59.05	$14.29^{o\|*\|*\|*}$	425.14
decision_tree	56.25	$37.50^{*\|o\|*\|o}$	397.56	69.52	$17.14^{o\|*\|*\|*}$	494.42

Table 3: QA accuracy after patching generated rationales with human feedback and regenerating answers. We show here the edit_acc, which is the percentage of examples for which the revised rationale successfully incorporates feedback and final_acc, which measures the final answer accuracy after regeneration with the revised rationale. Statistical significance at $p < 0.1$ is specified with $*$ (and o if not significant) with paired bootstrap tests in the order of the remaining rows in the table.

whether the final answer using the revised rationale is correct. We exclude all instances where the answer was leaked in feedback.

5.5 Results

We first show the results on standard decomposed QA for all rationale formats as well as standard answer generation (without rationales) on both Quoref and PubMedQA in Table 2. These results show that decomposed QA models can be competitive with end-to-end models. Although they slightly underperform standard answer generation for Quoref, decomposed QA is better performing on PubMedQA. This suggests that decomposed QA is a promising modeling approach, while being predisposed to provide more faithful rationales.

Annotator labels of sufficiency and faithfulness (presented in Figure 7) indicate that annotated_report and subquestions have rationales that are most often sufficient for both datasets, while markup_mask tends to lack most with sufficiency. On the other hand, extractive rationales from markup_mask are labeled most faithful for Quoref (58%), while annotated_report is relatively faithful for both datasets.

Our main results for repairing rationales through feedback are in Table 3. For Quoref, we find that rationale formats that expose the reasoning of the model are easier to repair through feedback. Interestingly, stricter formats with well-defined operations (such as procedural) are fairly effective for Quoref. On the other hand, for PubMedQA, rationale formats with more free-text components (such as annotated_report) that can allow more flexible edits are most effective. This is likely because comprehending medical articles and making inferences based on them can involve nuances that are harder to express with strict rationale formats. For example, feedback written for one PubMedQA example (Table 5) mentioned that the rationale didn't consider the fact that the study did not consider a control group for testing their hypothesis. This is easily incorporated into the free-text nature of the annotated_report, but is harder to incorporate in the procedural format.

We also note that when feedback is used to revise rationales that contain attributions, rationales can sometimes misquote sentences from the passage by hallucinating information that does not exist in the context. Although these revisions may result in correct answers, the rationales would be unfaithful, potentially decreasing user trust in the model.

In terms of annotator ease of providing feedback, we find that markup_mask is easiest to provide feedback for, because it may be easy to verbalize when information is missing from the rationale. However, these judgements do not correlate with actual effectiveness of the feedback for rationale repair. This suggests annotator ease of providing feedback may not correlate with actual effectiveness of feedback.

6 Study 2: Evaluating User Perception of Rationales

6.1 Setup

In the next study, we measure the extent to which different rationale formats enable users to *understand* and *trust* model responses. We sample examples where all 5 rationale formats have corresponding *correct answers* for both datasets. 40 annotators completed this study for Quoref examples while 44 completed it for PubMedQA examples.

6.2 Task

We collect Likert ratings of interpretability and trustworthiness of rationales. In addition, we col-

Q: What is the first name of the person whose son was a was a bachelor diplomat?				
Markup-and-mask	**Annotated Report**	**Procedural**	**Subquestions**	**Decision Tree**
Error location: Step 1 **Issue**: The information is insufficient to answer the question and the inference drawn from the context is incorrect. **Description**: Charles is the name of the son, and the question asks about the first name of the person who is the son's parent. **Suggestion**: The rationale needs to find the name of the son and then look for the name of the son's parent in the preceding context.	**Error location**: Step 1 **Issue**: The information is insufficient and the inference drawn from the context is incorrect. **Description**: The quote and annotation in step 1 reveal who the son is, whereas the question is asking about the first name of the parent, not the son. **Suggestion**: The rationale needs to find out who the son's parent is before providing their first name.	**Error location**: Step 1 **Issue**: Insufficient information. **Description**: Charles is the son who was a bachelor diplomat, and the question asks about the first name of Charles' parent. **Suggestion**: The rationale needs to locate who Charles' parent is in the text and then provide their first name.	**Error location**: Q1 **Issue**: The inference drawn from the context is incorrect. **Description**: The question asks about the first name of the parent mentioned in the passage as having a son who was a bachelor diplomat is not Charles Spencer Cowper, and not the first name of the bachelor diplomat himself. **Suggestion**: The rationale needs to find the bachelor diplomat's parent by looking at the context from the preceding sentences and then provide the parent's first name.	**Error location**: Q2-Yes **Issue**: The inference drawn from the context is incorrect. **Description**: Charles Spencer Cowper is the bachelor diplomat, and the question asks about the first name of the person Charles' parent. **Suggestion**: The rationale needs to look at previous sentences for the context of whose son Charles Spencer Cowper is and then provide that parent's first name.

Table 4: Examples of feedback collected for different rationale formats for Quoref examples in study 1 (§5).

Q: Is vitamin D insufficiency or deficiency related to the development of osteochondritis dissecans?				
Markup-and-mask	**Annotated Report**	**Procedural**	**Subquestions**	**Decision Tree**
Error location: Step 1 **Issue**: Insufficient information **Description**: There is not enough information to know whether vitamin D deficiency is related to the development of OCD lesions. Vitamin D could just be deficient in this population, and thus there could be many people with vitamin D deficiencies who never develop OCD lesions. **Suggestion**: The rationale needs to consider the presence of a control group. This could be vitamin D levels before developing an OCD lesion and/or vitamin D levels from a group of people who never developed OCD lesions.	**Error location**: Step 3 **Issue**: Incorrect inferences drawn from Context **Description**: Just because vitamin D levels are depleted amongst a group of OCD lesion patients does not mean that low vitamin D plays a role in the development of those lesions. For example, ... **Suggestion**: The rationale needs to consider the presence of a control group. This could be a measurement of vitamin D levels before, during, and after developing OCD lesions.	**Error location**: Step 1 **Issue**: Insufficient information **Description**: There is not a control group to compare the OCD patients' vitamin D levels to. Without a control group, we cannot know if Vitamin D is related to the development of OCD lesions. **Suggestion**: The rationale needs to consider the presence of a control group. Whether the researchers measured Vitamin D levels and OCD prevalence in the general population.	**Error location**: Q4 **Issue**: Insufficient information **Description**: The rationale says that the results suggest that low Vitamin D plays a role in the development of OCD lesions because vitamin D levels were depressed in a majority of the patients with OCD lesions. However, we do not have a control group/measurements and so cannot infer causality. **Suggestion**: The rationale needs to consider the presence of a control group. This could be pre-OCD lesion Vitamin D levels in the same set of subjects.	**Error location**: Q3-Yes **Issue**: Incorrect inferences drawn from the context **Description**: The model must have made an incorrect inference which caused them to not take the correct route down the decision tree and thus arrive at an incorrect answer. **Suggestion**: Considering whether a control group was included would have allowed us to better understand any causality between vitamin D levels and developing OCD lesions.

Table 5: Examples of feedback collected for different rationale formats for PubMedQA examples in study 1 (§5).

lect scalar judgements of the importance of different characteristics of rationales for annotators. Descriptions of Likert-scale options are in Figure 6.

Interpretability. A rationale should facilitate in making the model's reasoning more transparent to an end user. This is measured by asking annotators how beneficial the rationale is in helping them understand the reasoning process followed by the model. It is elicited on a scale of (*Very beneficial*, *A bit beneficial*, *Not beneficial at all*).

Trustworthiness. In addition to improving user understanding, a rationale that makes a model's decision-making transparent should do so in a way that helps users trust model responses. We ask annotators how likely they are to trust the model's answer, if the rationale was provided along with the answer. The rating is elicited on a scale of (*Very likely*, *A bit likely*, *A bit unlikely*, *Not likely at all*).

Scalar judgements. Next, we directly ask annotators for characteristics they value in rationales. They rate the following predefined rationale properties on a scale of 1-5 based on their importance:

- *Attribution*: Includes quotes from the context.
- *Depth of reasoning*: Provides detailed insight into the reasoning process.
- *Sequential reasoning*: Organized in a step-by-step manner.
- *Strictness*: Contains well-defined steps, with strict input and output formats.
- *Conciseness*: Brief and to the point.

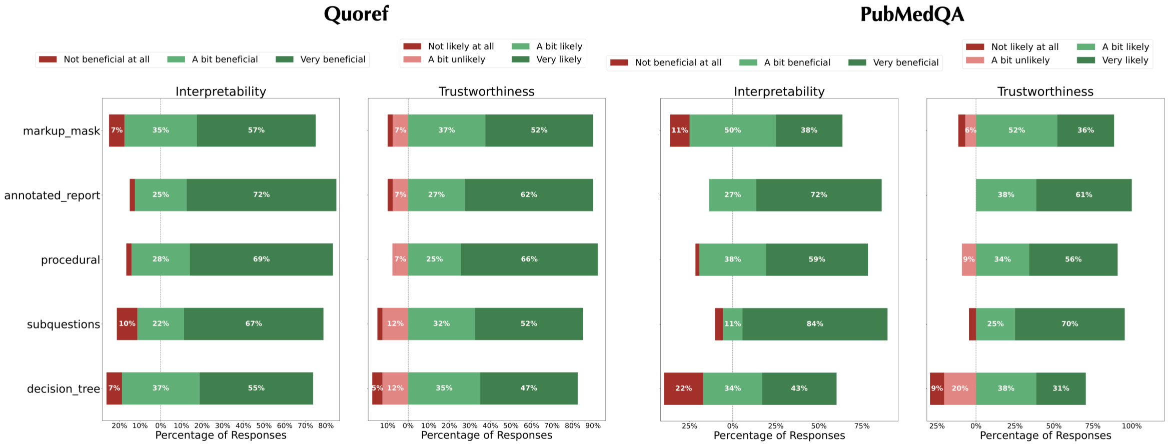


Figure 3: Likert distribution of the annotator judgements of interpretability & trustworthiness for different rationale formats corresponding to correct answers (§6).

6.3 Results

Figure 3 shows the Likert distributions of judgements of interpretability and trustworthiness for all formats on both datasets. These suggest that rationales with attributions and a sufficient amount of depth (`annotated_report` and `procedural`) are most easy to understand and trust for Quoref. On the other hand, `annotated_report` and `subquestions` rate highest on both axes for PubMedQA. Our interpretation of these judgements is that to be easily understandable and trustworthy for users, rationales should provide sufficient insight into the model's reasoning process and be accompanied with attributions.

Among the rationale properties presented to annotators, we find that attribution and depth of reasoning are rated as the most important properties of rationales. Figure 4 shows averaged scalar judgements for different rationale properties. A clear conclusion from these judgements is that providing attributions in the form of extracted quotes to the context is essential to users. This is likely because the attributions ground the model's reasoning into the context. In addition, depth of reasoning is highly valuable to users, especially for PubMedQA questions, where they may value a logical and coherent description of the model's reasoning.

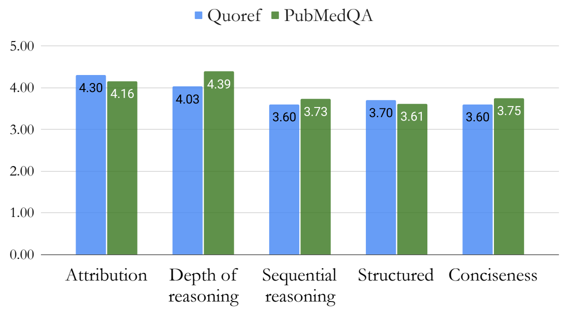


Figure 4: Scalar judgements of characteristics that annotators value in intermediate rationales (scale of 1-5).

7 Related Work

Decomposed QA. Although rationales from NLP models can be beneficial for users, there is recent evidence that shows that they are not always faithful to model responses (Ye and Durrett, 2022; Lyu et al., 2022; Turpin et al., 2023; Lanham et al., 2023). Decomposed question answering systems break down the QA problem into two stages, that of generating an intermediate rationale, and then using only that rationale to generate the answer (Lei et al., 2016; Eisenstein et al., 2022; Radhakrishnan et al., 2023). This provides a stronger inductive bias to the model to be faithful to the rationale. Similar ideas have been pursued in other tasks such as object recognition (Koh et al., 2020), image classification and text classification (Yeh et al., 2020). The precise format of the intermediate rationale that is optimal for human critiquability and interpretability is understudied. Our study is dedicated towards investigating the structure of this intermediate rationale.

Human feedback in NLP. Providing human feedback to NLP models has proven to be an effective way to repair models and fix model behaviors (Fernandes et al., 2023). Feedback can allow users to convey example-level critiques about model predictions, which, when incorporated into models, encourage them to perform better. Prior work has explored using human feedback for improving text summarization (Stiennon et al., 2020; Liu et al., 2023; Scheurer et al., 2023), question answering (Gao et al., 2022; Li et al., 2022), semantic parsing (Iyer et al., 2017; Elgohary et al., 2020, 2021), dialog generation (Shi et al., 2022; Ouyang et al., 2022; Xu et al., 2023), machine translation (Kreutzer et al., 2018) and image captioning (Fidler et al., 2017). Our work builds upon this prior work and investigates the effectiveness of human feedback for rationales provided by QA systems.

Rationales and explanations for NLP models. There is a large body of prior work studying explanations to supplement outputs from NLP models, both for improving models (Hancock et al., 2018; Lampinen et al., 2022; Wang et al., 2022; Zelikman et al., 2022; Zhou et al., 2023b) and explaining model outputs to end users. Prior work has found that explanations can be beneficial to end users for understanding model responses (DeYoung et al., 2020; Narang et al., 2020; Wiegreffe et al., 2022) as well as debugging models (Lertvittayakumjorn and Toni, 2021; Lamm et al., 2021). Prior work has also studied the impact of explanations on user trust (Papenmeier et al., 2022) and performance of human-AI teams (Bansal et al., 2021). Boyd-Graber et al. (2022); Jacovi and Goldberg (2020) provide useful guidelines to conduct human-centered and faithful evaluations of these explanations. We conduct another such evaluation that is centered on the format of model rationales presented to end users.

8 Conclusion

Our work analyzed how model-generated explanations or rationales should be formatted to be most amenable to repair through user feedback. We also collected qualitative judgements of how different formats enable users to understand and trust model outputs. We found that rationale formats significantly affect what rationales are amenable to be repaired through feedback. In terms of user perception of rationales, we find that some rationale formats, such as the *annotated report*, are more favorable for enabling users to understand and trust model responses. Finally, we find that among a few properties considered, attribution and depth of reasoning are the most important characteristics of rationales to users. We hope that this work can help researchers and practitioners alike make informed decisions about how to present language model responses and collect feedback from end users.

9 Limitations

Rationale Formats. The rationale formats we consider are by no means exhaustive and there could be numerous other plausible formats for intermediate rationales. We choose a set of rationales that vary according to some well-defined properties (mentioned in Table 1), that can allow us to form conclusions about the importance of those properties.

Feedback Structure. We choose a feedback structure that encompasses a few crucial aspects of feedback highlighted in previous work. However, it may be possible that there are other types of feedback that show different trends in effectiveness across rationale formats.

Scope of QA problems. We choose reading comprehension datasets where questions are formulated based on a given context. While these may not be representative of all forms of QA problems, we hope our findings can broadly inform practitioners about ways to present QA system responses to users (for instance, when deploying retrieval-augmented QA systems).

Acknowledgements

First, we would like to thank the annotators who participated in our studies. We would also like to thank Elizabeth Sieber, Jacob Eisenstein, Weiqiu You and Eric Wong for helpful comments and discussions.

References

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–16.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32, Seattle, United States. Association for Computational Linguistics.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. 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.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Eisenstein, Daniel Andor, Bernd Bohnet, Michael Collins, and David Mimno. 2022. Honest students from untrusted teachers: Learning an interpretable question-answering pipeline from a pretrained language model. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

Ahmed Elgohary, Saghar Hosseini, and Ahmed Hassan Awadallah. 2020. Speak to your parser: Interactive text-to-SQL with natural language feedback. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2065–2077, Online. Association for Computational Linguistics.

Ahmed Elgohary, Christopher Meek, Matthew Richardson, Adam Fourney, Gonzalo Ramos, and Ahmed Hassan Awadallah. 2021. NL-EDIT: Correcting semantic parse errors through natural language interaction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5599–5610, Online. Association for Computational Linguistics.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *arXiv preprint arXiv:2305.00955*.

Sanja Fidler et al. 2017. Teaching machines to describe images with natural language feedback. *Advances in Neural Information Processing Systems*, 30.

Ge Gao, Eunsol Choi, and Yoav Artzi. 2022. Simulating bandit learning from user feedback for extractive question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5167–5179, Dublin, Ireland. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973, Vancouver, Canada. 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.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1264–1279, Online. Association for Computational Linguistics.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy

Liang. 2020. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR.

Julia Kreutzer, Shahram Khadivi, Evgeny Matusov, and Stefan Riezler. 2018. Can neural machine translation be improved with user feedback? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 92–105, New Orleans - Louisiana. Association for Computational Linguistics.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

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.

Piyawat Lertvittayakumjorn and Francesca Toni. 2021. Explanation-based human debugging of nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 9:1508–1528.

Zichao Li, Prakhar Sharma, Xing Han Lu, Jackie Cheung, and Siva Reddy. 2022. Using interactive feedback to improve the accuracy and explainability of question answering systems post-deployment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 926–937, Dublin, Ireland. Association for Computational Linguistics.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2022. Can large language models reason about medical questions? *arXiv preprint arXiv:2207.08143*.

Yixin Liu, Budhaditya Deb, Milagro Teruel, Aaron Halfaker, Dragomir Radev, and Ahmed Hassan Awadallah. 2023. On improving summarization factual consistency from natural language feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15144–15161, Toronto, Canada. Association for Computational Linguistics.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in nlp: A survey. *arXiv preprint arXiv:2209.11326*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Laura Martignon, Oliver Vitouch, Masanori Takezawa, and Malcolm R Forster. 2003. Naive and yet enlightened: From natural frequencies to fast and frugal decision trees. *Thinking: Psychological perspectives on reasoning, judgment and decision making*, pages 189–211.

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 preprint arXiv:2004.14546*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*.

Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's complicated: The relationship between user trust, model accuracy and explanations in ai. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4):1–33.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2023. Training language models with language feedback at scale. *arXiv preprint arXiv:2303.16755*.

Weiyan Shi, Emily Dinan, Kurt Shuster, Jason Weston, and Jing Xu. 2022. When life gives you lemons, make cherryade: Converting feedback from bad responses into good labels. *arXiv preprint arXiv:2210.15893*.

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. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Simeng Sun, Yang Liu, Shuohang Wang, Chenguang Zhu, and Mohit Iyyer. 2023. Pearl: Prompting large language models to plan and execute actions over long documents. *arXiv preprint arXiv:2305.14564*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Rationale-augmented ensembles in language models. *arXiv preprint arXiv:2207.00747*.

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.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2023. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2023a. Least-to-most prompting enables complex reasoning in large language models. *ICLR*.

Yangqiaoyu Zhou, Yiming Zhang, and Chenhao Tan. 2023b. Flame: Few-shot learning from natural language explanations. *arXiv preprint arXiv:2306.08042*.

A Experimental Details

Prompts. The prompts used to generate rationales for all formats are provided in Tables 7, 8, 9, 10 and 11, while the prompts used to generate answers are provided Tables 12, 13, 14, 15 and 16. For generating answers for PubMedQA, we modify the prompt same way as previous work (Liévin et al., 2022), by transforming it into a multiple-choice question. For revising rationales, we use a similar format as these prompts but also includes the following string in the instruction –

`Correct the given Rationale based on the Feedback. The Feedback first points out the Error Location, then mentions the Issue and gives a Description of the issue, and finally provides a Suggestion to correct the given Rationale. The Rationale is required to be sufficient to answer the Question on its own and faithful to the Context..`

Hyperparameter settings. Rationales and answers were sampled from `gpt-3.5-turbo` with a temperature of 0.0 and a maximum length of 512 tokens when sampling rationales, and 64 tokens when sampling answers. For prompting models, we sample few-shot exemplars using BM25 (Robertson et al., 2009) from a set of 100 manually labeled examples with rationales. We sample as many exemplars as can fit within the maximum sequence length (4096) of the model.

B Annotation Details

Annotator backgrounds. For both studies, annotators were recruited from Prolific[3], and required to be fluent in English. They were required to have at least 100 accepted submissions and an approval rate of at least 99%. They were also required to have at least a bachelor's degree.

Annotators for the Quoref task were based in UK, USA, Australia, Ireland, Canada or New Zealand. Annotators for the PubMedQA task were based in UK, USA, Ireland, Germany, France, Australia, Canada, Denmark, Netherlands, Switzerland, Norway, Portugal or Sweden. These annotators were additionally required to be employed in the healthcare/medicine sector.

Annotation costs. In both studies, annotators were compensated at the rate of $15 per hour with additional bonuses when annotators spent more time than we anticipated.

Annotation interface. Figures 5 and 6 show screenshots of our annotation interface for both Study 1 and Study 2 in the order the task was presented to annotators.

C Additional Results

Table 6 shows the effectiveness of feedback with examples where 3 rationale formats get the answer wrong. Figure 7 shows the Likert distribution of sufficiency, faithfulness and ease of providing feedback for all rationale formats for both datasets.

[3] `www.prolific.co`

Rationale Format	Quoref			PubMedQA		
	edit_acc	final_acc	time_taken	edit_acc	final_acc	time_taken
markup_mask	70.37	62.96	340.29	64.58	33.33	389.79
annotated_report	61.29	58.06	290.17	58.06	**45.16**	447.90
procedural	52.94	58.82	345.55	51.28	33.33	333.80
subquestions	81.81	**72.72**	316.71	65.12	30.23	348.57
decision_tree	66.67	38.10	340.39	88.64	13.64	465.39

Table 6: Results after patching generated rationales with human feedback for examples where the answer is wrong for 3 rationale formats. We show here the `edit_acc`, which measures if the revised rationale successfully incorporates feedback and `final_acc`, which measures the final accuracy after regeneration with the revised rationale.

X2R Prompt (markup_mask)

```
Extract the most relevant 1-2 sentences from the context as a rationale
sufficient to answer the question. Also resolve any ambiguous terms and
coreferences in the extracted sentences to make them standalone. The relevant
sentences should be sufficient to determine the answer to the question.

Context: [CONTEXT]

Question: [QUESTION]

Rationale:
```

Table 7: $X2R$ prompt for the `markup_mask` format.

X2R Prompt (annotated_report)

```
Generate a rationale that is helpful and sufficient to answer the question. The
rationale should contain a list of extracted phrases from the context and the
conclusion drawn from each phrase. Try to include no more than 5 extracted
phrases.

Context: [CONTEXT]

Question: [QUESTION]

Annotations:
```

Table 8: $X2R$ prompt for the `annotated_report` format.

X2R Prompt (`procedural`)
Construct a structured Plan for answering the Question, that should provide a sequential process for finding the answer. The Plan should not directly answer the Question but only provide the reasoning. You can use the following operations in the plan: - Extract-relevant-sentences: Extract relevant sentences from the passage that are sufficient to answer the question. The extracted sentences should include the necessary information to answer the question accurately. - Disambiguate-question-entity(s): Determine the specific entity or phrase that the string s in the question refers to. Clarify any ambiguous terms or references to ensure a precise understanding. - Disambiguate-plan-entity(s): Identify the entity or phrase that the string s in the plan refers to. Resolve any ambiguity within the plan by specifying the relevant entities explicitly. - Locate-entity(s): Generate a subquestion s that is important to answer the original question without simply repeating the original question. Determine the exact entity or phrase that provides the answer to the subquestion s. Context: [CONTEXT] Question: [QUESTION] Plan:

Table 9: $X2R$ prompt for the `procedural` format.

X2R Prompt (`subquestions`)
Form subquestions required to answer the given question based on the passage. You cannot repeat the given question as a subquestion. The formed subquestions and their answers should be sufficient to answer the given question. Try to form no more than 5 subquestions. Context: [CONTEXT] Question: [QUESTION] Subquestions:

Table 10: $X2R$ prompt for the `subquestions` format.

X2R Prompt (decision_tree)

```
Generate a decision tree-based rationale to answer the question. The decision
tree should be sufficient to answer the question. However, it should not answer
the question directly. Try to form no more than 5 subquestions.

Context: [CONTEXT]

Question: [QUESTION]

Decision Tree:
```

Table 11: $X2R$ prompt for the `decision_tree` format.

R2Y Prompt (markup_mask)

```
Use these extracted relevant sentences from a passage to answer the question.

Relevant sentences: [RATIONALE]

Question: [QUESTION]

Answer:
```

Table 12: $R2Y$ prompt for the `markup_mask` format.

R2Y Prompt (annotated_report)

```
You are given an annotated rationale from a passage as context. The annotations
are in the format of a list of extracted phrases from the context and the
conclusion drawn from each phrase. Answer the question based on the rationale
alone.

Rationale: [RATIONALE]

Question: [QUESTION]

Answer:
```

Table 13: $R2Y$ prompt for the `annotated_report` format.

R2Y Prompt (`procedural`)

`Answer the Question based on the Plan-based Rationale. The Plan gives a sequential process of finding the answer. The following operations can be used in a plan: <Skipped for brevity>.`
 	
`Plan: [RATIONALE]`
 	
`Question: [QUESTION]`
 	
`Answer:`

Table 14: $R2Y$ prompt for the `procedural` format.

R2Y Prompt (`subquestions`)

`Answer the given Question solely based on the Subquestions and their answers. The answer can always be found from the Subquestions so make your best guess.`
 	
`Subquestions: [RATIONALE]`
 	
`Question: [QUESTION]`
 	
`Answer:`

Table 15: $R2Y$ prompt for the `subquestions` format.

R2Y Prompt (`decision_tree`)

`Answer the Question solely based on the Decision Tree-based Rationale.`
 	
`Decision Tree: [RATIONALE]`
 	
`Question: [QUESTION]`
 	
`Answer:`

Table 16: $R2Y$ prompt for the `decision_tree` format.

1. (Study 1 + Study 2) Context + Question + Rationale + Generated Correct Answer

Context

BACKGROUND. Implant-related infections represent one of the most severe complications in orthopaedics. A fast-resorbable, antibacterial-loaded hydrogel may reduce or prevent bacterial colonization and biofilm formation of implanted biomaterials.QUESTIONS/

PURPOSES. We asked: (1) Is a fast-resorbable hydrogel able to deliver antibacterial compounds in vitro? (2) Can a hydrogel (alone or antibacterial-loaded) coating on implants reduce bacterial colonization? And (3) is intraoperative coating feasible and resistant to press-fit implant insertion?

METHODS. We tested the ability of Disposable Antibacterial Coating (DAC) hydrogel (Novagenit Srl, Mezzolombardo, Italy) to deliver antibacterial agents using spectrophotometry and a microbiologic assay. Antibacterial and antibiofilm activity were determined by broth microdilution and a crystal violet assay, respectively. Coating resistance to press-fit insertion was tested in rabbit tibias and human femurs.

RESULTS. Complete release of all tested antibacterial compounds was observed in less than 96 hours. Bactericidal and antibiofilm effect of DAC hydrogel in combination with various antibacterials was shown in vitro. Approximately 80% of the hydrogel coating was retrieved on the implant after press-fit insertion.

Question

Does implant coating with antibacterial-loaded hydrogel reduce bacterial colonization and biofilm formation in vitro?

Rationale (Format: Decision Tree)

- **Start**
- **--> Q1: Did the study test the ability of a fast-resorbable hydrogel to deliver antibacterial compounds in vitro?**
  - **--> Yes:**
    - **--> Q2: Was the release of all tested antibacterial compounds observed in less than 96 hours?**
      - **--> Yes:**
        - **--> Q3: Was the bactericidal and antibiofilm effect of the hydrogel in combination with various antibacterials demonstrated in vitro?**
          - **--> Yes:**
            - **--> Outcome:** Implant coating with antibacterial-loaded hydrogel may reduce bacterial colonization and biofilm formation in vitro
          - **--> No:**
            - **--> Outcome:** The effectiveness of implant coating with antibacterial-loaded hydrogel in reducing bacterial colonization and biofilm formation in vitro is unclear
      - **--> No:**
        - **--> Outcome:** Cannot evaluate the effectiveness of implant coating with antibacterial-loaded hydrogel in reducing bacterial colonization and biofilm formation in vitro
  - **--> No:**
    - **--> Outcome:** Cannot evaluate the effectiveness of implant coating with antibacterial-loaded hydrogel in reducing bacterial colonization and biofilm formation in vitro
- **End**

Generated Correct Answer

yes

(Study 1) Generated Incorrect Answer is also added

Generated Incorrect Answer

no

(Study 1 + Study 2) Different examples of rationales that could appear above (in addition to decision tree above)

Rationale (Format: Annotated Report)

1. Quote from context: "Is a fast resorbable hydrogel able to deliver antibacterial compounds in vitro?"
Annotation: The first question being asked in the study, indicating the purpose of examining the hydrogel's ability to deliver antibacterial compounds.

2. Quote from context: "Bactericidal and antibiofilm effect of DAC hydrogel in combination with various antibacterials was shown in vitro"
Annotation: Indicates that the hydrogel coating, when combined with antibacterials, demonstrated effectiveness in killing bacteria and preventing biofilm formation in a laboratory setting.

3. Quote from context: "Approximately 80% of the hydrogel coating was retrieved on the implant after press fit insertion"
Annotation: Shows that the hydrogel coating remained intact on the implant even after being subjected to press-fit insertion, suggesting it can be used effectively in real-world scenarios.

Rationale (Format: Markup and Mask)

1. Quote from context: "Bactericidal and antibiofilm effect of DAC **[Disposable Antibacterial Coating]** hydrogel in combination with various antibacterials was shown in vitro."

Rationale (Format: Procedural)

1) **Extract-relevant-sentences:** "We tested the ability of Disposable Antibacterial Coating (DAC) hydrogel (Novagenit Srl, Mezzolombardo, Italy) to deliver antibacterial agents using spectrophotometry and a microbiologic assay.", "Bactericidal and antibiofilm effect of DAC hydrogel in combination with various antibacterials was shown in vitro."

2) **Disambiguate-question-entity ("implant coating"):** "implantation of hydrogel on the surface of an implant".

3) **Disambiguate-plan-entity ("hydrogel"):** "Disposable Antibacterial Coating (DAC) hydrogel".

4) **Locate-entity ("what is the effect of hydrogel coating on bacterial colonization and biofilm formation?"):** "reduction or prevention of both bacterial colonization and biofilm formation".

5) **Locate-entity ("where was the effect of hydrogel coating tested?"):** "in vitro".

Rationale (Format: Subquestions)

Q1: What was the purpose of the study?

A: The purpose of the study was to assess the ability of a fast-resorbable hydrogel to deliver antibacterial compounds in vitro and determine if hydrogel coating on implants can reduce bacterial colonization.

Q2: What testing methods were used to assess the antibacterial and antibiofilm activity of the hydrogel?

A: Bactericidal and antibiofilm activity were determined using broth microdilution and a crystal violet assay.

Q3: Was the hydrogel coating able to deliver antibacterial compounds effectively?

A: Yes, complete release of all tested antibacterial compounds was observed in less than 96 hours.

Q4: Did the hydrogel coating reduce bacterial colonization and biofilm formation in vitro?

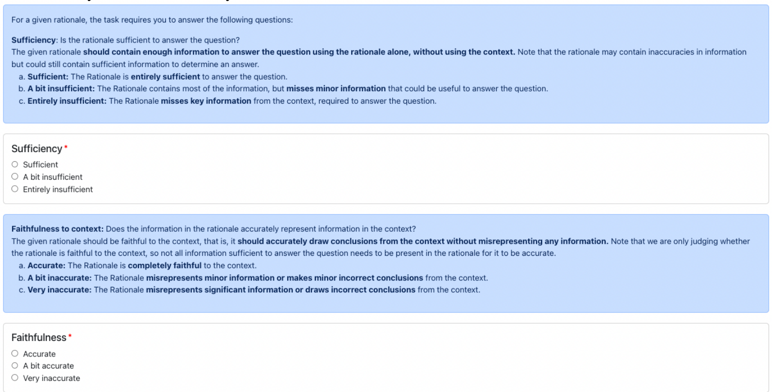
A: Yes, the combination of the hydrogel coating with various antibacterials showed bactericidal and antibiofilm effects in vitro.

Q5: Was the hydrogel coating resistant to press-fit implant insertion?

A: Approximately 80% of the hydrogel coating was retrieved on the implant after press-fit insertion, indicating resistance to the insertion process.

Figure 5: Screenshots of the interface (1).

2. (Study 1) Sufficiency + Faithfulness



For a given rationale, the task requires you to answer the following questions:

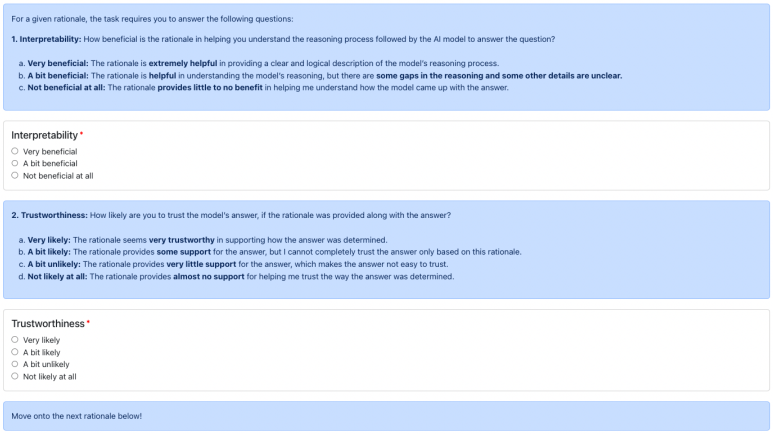
Sufficiency: Is the rationale sufficient to answer the question?
The given rationale **should contain enough information to answer the question using the rationale alone, without using the context.** Note that the rationale may contain inaccuracies in information but could still contain sufficient information to determine an answer.
a. **Sufficient:** The Rationale is **entirely sufficient** to answer the question.
b. **A bit insufficient:** The Rationale contains most of the information, but **misses minor information** that could be useful to answer the question.
c. **Entirely insufficient:** The Rationale **misses key information** from the context, required to answer the question.

Sufficiency *
- Sufficient
- A bit insufficient
- Entirely insufficient

Faithfulness to context: Does the information in the rationale accurately represent information in the context?
The given rationale should be faithful to the context, that is, it **should accurately draw conclusions from the context without misrepresenting any information.** Note that we are only judging whether the rationale is faithful to the context, so not all information sufficient to answer the question needs to be present in the rationale for it to be accurate.
a. **Accurate:** The Rationale is **completely faithful** to the context.
b. **A bit inaccurate:** The Rationale **misrepresents minor information or makes minor incorrect conclusions** from the context.
c. **Very inaccurate:** The Rationale **misrepresents significant information or draws incorrect conclusions** from the context.

Faithfulness *
- Accurate
- A bit accurate
- Very inaccurate

2. (Study 2) Interpretability + Trustworthiness



For a given rationale, the task requires you to answer the following questions:

1. Interpretability: How beneficial is the rationale in helping you understand the reasoning process followed by the AI model to answer the question?

a. **Very beneficial:** The rationale is **extremely helpful** in providing a clear and logical description of the model's reasoning process.
b. **A bit beneficial:** The rationale is **helpful** in understanding the model's reasoning, but there are **some gaps in the reasoning and some other details are unclear.**
c. **Not beneficial at all:** The rationale **provides little to no benefit** in helping me understand how the model came up with the answer.

Interpretability *
- Very beneficial
- A bit beneficial
- Not beneficial at all

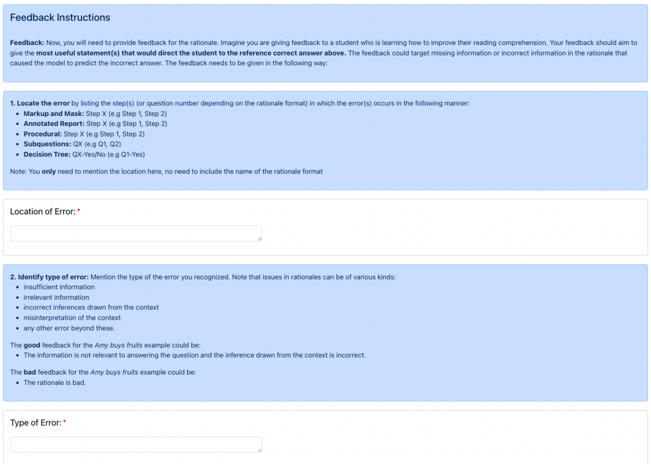
2. Trustworthiness: How likely are you to trust the model's answer, if the rationale was provided along with the answer?

a. **Very likely:** The rationale seems **very trustworthy** in supporting how the answer was determined.
b. **A bit likely:** The rationale provides **some support** for the answer, but I cannot completely trust the answer only based on this rationale.
c. **A bit unlikely:** The rationale provides **very little support** for the answer, which makes the answer not easy to trust.
d. **Not likely at all:** The rationale provides **almost no support** for helping me trust the way the answer was determined.

Trustworthiness *
- Very likely
- A bit likely
- A bit unlikely
- Not likely at all

Move onto the next rationale below!

3. (Study 1) Feedback: Instructions + Location of Error + Type of Error + Description of Error + Actionable Suggestion



Feedback Instructions

Feedback: Now, you will need to provide feedback for the rationale. Imagine you are giving feedback to a student who is learning how to improve their reading comprehension. Your feedback should aim to give the **most useful statement(s) that would direct the student to the reference correct answer above.** The feedback could target missing information or incorrect information in the rationale that caused the model to predict the incorrect answer. The feedback needs to be given in the following way:

1. Locate the error by listing the step(s) (or question number depending on the rationale format) in which the error(s) occurs in the following manner:
- **Markup and Mask:** Step X (e.g Step 1, Step 2)
- **Annotated Report:** Step X (e.g Step 1, Step 2)
- **Procedural:** Step X (e.g Step 1, Step 2)
- **Subquestions:** QX (e.g Q1, Q2)
- **Decision Tree:** QX-Yes/No (e.g Q1-Yes)

Note: You **only** need to mention the location here, no need to include the name of the rationale format

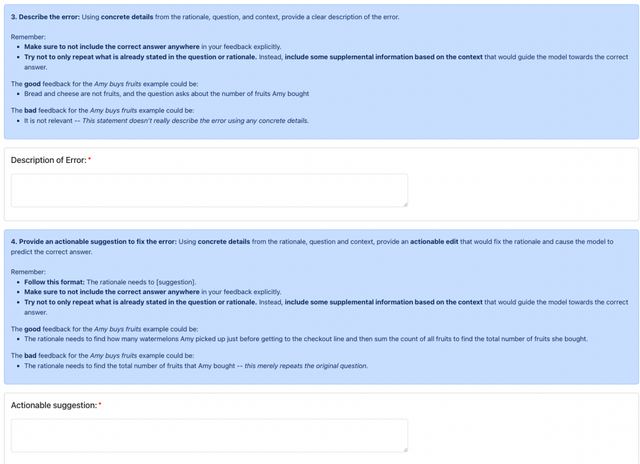
Location of Error: *

2. Identify type of error: Mention the type of the error you recognized. Note that issues in rationales can be of various kinds:
- insufficient information
- irrelevant information
- incorrect inferences drawn from the context
- misinterpretation of the context
- any other error beyond these.

The **good** feedback for the *Amy buys fruits* example could be:
- The information is not relevant to answering the question and the inference drawn from the context is incorrect.

The **bad** feedback for the *Amy buys fruits* example could be:
- The rationale is bad.

Type of Error: *



3. Describe the error: Using **concrete details** from the rationale, question, and context, provide a clear description of the error.

Remember:
- **Make sure to not include the correct answer anywhere** in your feedback explicitly.
- **Try not to only repeat what is already stated in the question or rationale.** Instead, **include some supplemental information based on the context** that would guide the model towards the correct answer.

The **good** feedback for the *Amy buys fruits* example could be:
- Bread and cheese are not fruits, and the question asks about the number of fruits Amy bought

The **bad** feedback for the *Amy buys fruits* example could be:
- It is not relevant -- *This statement doesn't really describe the error using any concrete details.*

Description of Error: *

4. Provide an actionable suggestion to fix the error: Using **concrete details** from the rationale, question and context, provide an **actionable edit** that would fix the rationale and cause the model to predict the correct answer.

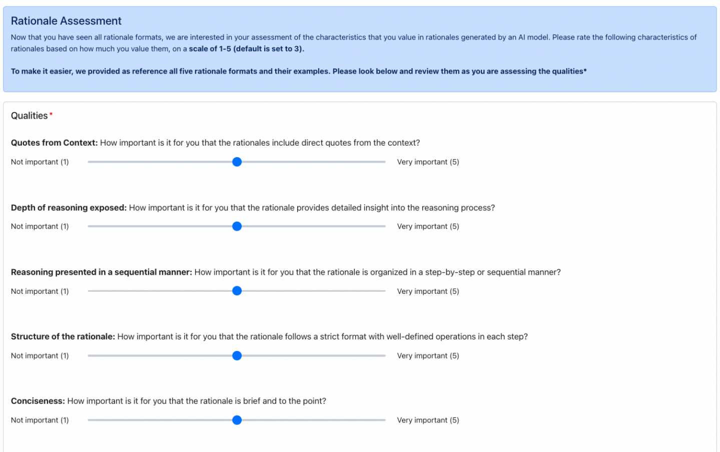
Remember:
- **Follow this format:** The rationale needs to [suggestion].
- **Make sure to not include the correct answer anywhere** in your feedback explicitly.
- **Try not to only repeat what is already stated in the question or rationale.** Instead, **include some supplemental information based on the context** that would guide the model towards the correct answer.

The **good** feedback for the *Amy buys fruits* example could be:
- The rationale needs to find how many watermelons Amy picked up just before getting to the checkout line and then sum the count of all fruits to find the total number of fruits she bought.

The **bad** feedback for the *Amy buys fruits* example could be:
- The rationale needs to find the total number of fruits that Amy bought -- *this merely repeats the original question*.

Actionable suggestion: *

3. (Study 2) Rationale Assessment



Rationale Assessment

Now that you have seen all rationale formats, we are interested in your assessment of the characteristics that you value in rationales generated by an AI model. Please rate the following characteristics of rationales based on how much you value them, on a **scale of 1-5 (default is set to 3).**

To make it easier, we provided as reference all five rationale formats and their examples. Please look below and review them as you are assessing the qualities*

Qualities *

Quotes from Context: How important is it for you that the rationales include direct quotes from the context?
Not important (1) Very important (5)

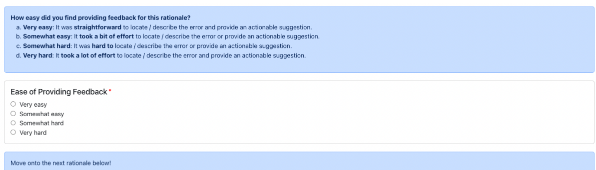
Depth of reasoning exposed: How important is it for you that the rationale provides detailed insight into the reasoning process?
Not important (1) Very important (5)

Reasoning presented in a sequential manner: How important is it for you that the rationale is organized in a step-by-step or sequential manner?
Not important (1) Very important (5)

Structure of the rationale: How important is it for you that the rationale follows a strict format with well-defined operations in each step?
Not important (1) Very important (5)

Conciseness: How important is it for you that the rationale is brief and to the point?
Not important (1) Very important (5)

4. (Study 1) Ease of Providing Feedback



How easy did you find providing feedback for this rationale?
a. **Very easy**: It was **straightforward** to locate / describe the error and provide an actionable suggestion.
b. **Somewhat easy**: It **took a bit of effort** to locate / describe the error or provide an actionable suggestion.
c. **Somewhat hard**: It was **hard to** locate / describe the error or provide an actionable suggestion.
d. **Very hard**: It **took a lot of effort** to locate / describe the error and provide an actionable suggestion.

Ease of Providing Feedback *
- Very easy
- Somewhat easy
- Somewhat hard
- Very hard

Move onto the next rationale below!

Figure 6: Screenshots of the interface (2-3).

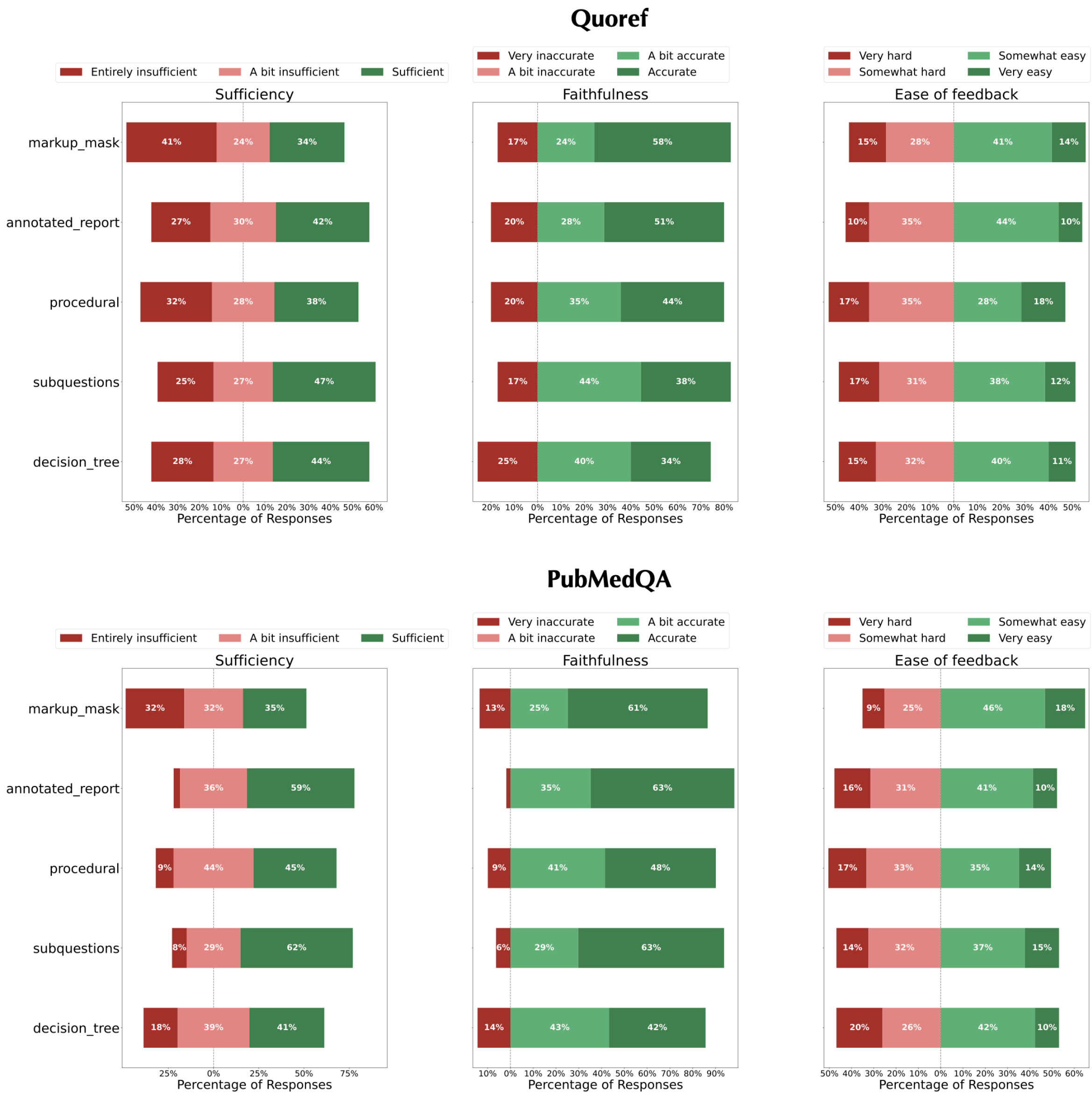


Figure 7: Likert distribution of the sufficiency & faithfulness for different rationale formats, as well as ease of writing feedback. (§5).