Characterizing Large Language Models as Rationalizers of Knowledge-intensive Tasks

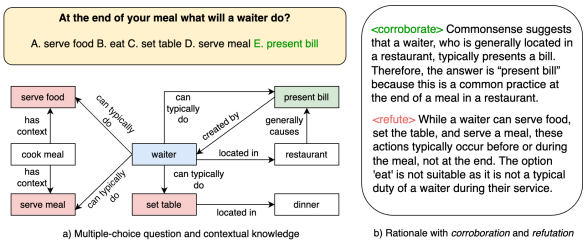
Anonymous ACL submission

Abstract

Large language models (LLMs) are proficient at generating fluent text with minimal task-specific supervision. However, their ability to generate rationales for knowledge-intensive tasks (KITs) remains under-explored. Generating rationales for KIT solutions, such as commonsense multiple-choice QA, requires external knowledge to support predictions and refute alternate options. In this work, we consider the task of generating retrieval-augmented rationalization of KIT model predictions via external knowledge-guidance within a few-shot setting. Surprisingly, crowd-workers preferred LLM-generated rationales over existing crowd-sourced rationales, generated in a similar knowledge-guided setting, on aspects such as factuality, sufficiency, and convincingness. However, follow-up fine-grained evaluation of such rationales highlight the need for further improvements in conciseness, novelty, and domain invariance. Additionally, through an expert-sourced study evaluating the reliability of the rationales, we demonstrate that humans' trust in LLM-generated rationales erode when communicated faithfully, *i.e.*, without taking model prediction accuracy into account. We find that even instrumenting simple guardrails can be an effective for reliable rationalization.

1 Introduction

In recent years, generating *rationales* (*i.e.*, free-text explanations) of natural language understanding tasks has been increasingly explored in the field of explainable NLP. Such rationales — while less functionally grounded, *i.e.*, they may not entirely reflect the model's behavior — provide an effective interface to interpretably communicate model decisions to end-users (Hendricks et al., 2016; Camburu et al., 2018; Madsen et al., 2022; Gurrapu et al., 2023). Generating these rationales via direct supervision (Ehsan et al., 2018; Narang et al., 2020) or fine-tuning (Aggarwal et al., 2021; Rei et al., 2022) requires the collection of high-quality human-authored rationales. Collecting such rationales via crowd-sourcing is expensive, difficult to standardize, and lacks generalizability to different domains (Wiegreffe and Marasović, 2021; Tan, 2021). Recent work (Wiegreffe et al., 2022) showcases that large language model (LLM) generated rationales, obtained via few-shot in-context learning (Radford et al., 2019; Brown et al., 2020; Huang et al., 2023), alleviate these challenges while showcasing surprising effectiveness over crowdsourced rationales on dimensions such as human preference. However, characterizing the suitability of LLMs as rationalizers of knowledge-intensive task (KIT) decisions such as commonsense question answering (CSQA (Talmor et al., 2019)) and open book question answering (OBQA (Mihaylov et al., 2018)) requires further investigation due to the difference in scope and setting from prior work (Wiegreffe et al., 2022).



a) Multiple-choice question and contextual knowledge

b) Rationale with *corroboration* and *refutation*

Figure 1: a) A commonsense question with multiple choices and knowledge extracted from ConceptNet and b) proposed LLM-generated rationale corroborating the selected answer and refuting the other choices.

Firstly, KITs such as CSQA and OBQA are framed as multiple-choice questions, requiring models to select one answer from several choices (see Figure 1a). Therefore, a corresponding well-formed rationale is required to be (a) comprehensive, *i.e.*, state facts that are not present in the question but are essential for rationalization, and (b) refutation complete, *i.e.*, rationalize why the rest of the choices are incorrect or not best suited as

the answer (Aggarwal et al., 2021). We show an example of such a rationale in Figure 1b. However, LLM-generated rationales in prior work (Wiegreffe et al., 2022) have only been evaluated on their corroboration capabilities. Secondly, LLM-generated rationales in prior work are abstractive (Gurrapu et al., 2023), lacking grounding on external knowledge sources crucial for accomplishing the task — KIT models designed for CSQA and OBQA (Feng et al., 2020; Yasunaga et al., 2021, 2022) refer to external sources such as ConceptNet (Speer et al., 2017) (see Figure 1a). Finally, KIT models may predict incorrectly — faithfully rationalizing such mistakes may erode the end-user's trust in the generated rationales. Existing approaches in explainable NLP omit the incorrect prediction confounder and evaluate only rationales of correct predictions. However, with LLM-generated rationales being increasingly adopted in real-world scenarios, such as rationalizing why a candidate is suitable for an advertised job[1], it is important to scrutinize the practical implications of such deployments and inform guidelines for safe and responsible adoption.

Given the setting of generating corroborating and refutation complete rationales of KIT model decisions, we explore the suitability of retrieval-augmented rationale generation using LLMs. We enrich the prompt to LLMs with relevant knowledge retrieved from external sources to condition the rationale generation on facts. More specifically, we generate knowledge-guided rationales containing corroboration and refutation components — similar to Figure 1b — via few-shot prompting of LLMs. We conducted three human subject studies to evaluate the effectiveness of such rationales in communicating KIT model decisions. The observations from these studies enable coarse- and fine-grained characterization of the strengths and weaknesses of LLM-generated knowledge-guided rationalization of KIT model decisions.

More specificlly, we conduct two studies via crowdsourcing to evaluate the preferability and acceptability of such rationales to crowd-workers. In another study involving experts — motivated by existing literature on trust in explainable AI (Hoffman et al., 2018; Stites et al., 2021) — we explore the implications of faithfully rationalizing KIT model decisions irrespective of their correctness. The crowd-sourced studies demonstrate that, more often than not, crowdworkers prefer LLM-generated rationales to crowdsourced rationales in existing datasets, citing their factuality, sufficiency, and convincing refutation. Follow-up fine-grained analysis reveals that LLM-generated rationales still have significant room for improvement along dimensions such as *insightfulness* (*i.e.*, providing new information), *redundancy* (*i.e.*, avoiding repetitive text), and *generalizability* (*i.e.*, domain invariance.) The expert-sourced study confirms that faithful rationalization of incorrect model predictions degrades humans' trust in the generated rationales. We further explore the utility of instrumenting mechanisms to intervene the incorrect predictions via a review-then-rationalize pipeline instead of faithfully rationalizing and find that even simple strategies may help intervene up to 71% of the incorrect predictions. We will publicly release the code and data.

2 Knowledge-enhanced Rationalization

KIT models such as MHGRN (Feng et al., 2020), QAGNN (Yasunaga et al., 2021), and Dragon (Yasunaga et al., 2022) combine language model and knowledge graph representations to solve complex tasks such as commonsense QA (Talmor et al., 2019). We aim to generate rationales that corroborate the KIT model's prediction with additional relevant facts while refuting the other choices (see Figure 1.) Our approach is similar to existing retrieval-augmented generation strategies with LLMs (Peng et al., 2023; Lazaridou et al., 2022; Zhao et al., 2023; Mei et al., 2023). To guide the generation of these rationale components, *i.e.*, corroboration and refutation, we retrieve facts concerning the knowledge-intensive task — *e.g.*, questions and choices in CSQA and OBQA — from a knowledge graph such as ConceptNet (Speer et al., 2017). We then prompt an LLM to rationalize the prediction via conditioning on the provided knowledge. Figure 2 outlines the rationalization process given an *input*, *i.e.*, question, choices, and model prediction.

Given an external knowledge-graph such as ConceptNet (Speer et al., 2017), we employ the knowledge extraction strategy used in QAGNN (Yasunaga et al., 2021) to first retrieve the facts relevant to a question and then select top-k ($k = 5$) facts based on their RoBERTa (Liu et al., 2019) score given the question and a choice. Such selection enables us to fit the knowledge facts within the token limits of an LLM prompt. We employ greedy decoding-based few-shot prompting to query an

[1] https://www.businessinsider.com/sc/indeed-is-embracing-ai-to-power-the-future-of-work

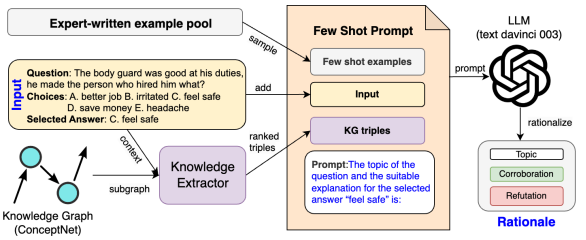


Figure 2: Given an Input (*i.e.*, QA and model prediction), an LLM is prompted to generate a rationale with few-shot examples sampled from an expert-written pool.

LLM for rationalization. Each example in the prompt contains a QA task, the corresponding KIT model prediction, facts retrieved from ConceptNet, and expert written rationale corrborating the prediction and refuting other choices. We opted for expert-authored rationales due to their reported effectiveness over crowdsourced rationales (Wiegreffe et al., 2022). The paper's authors collaboratively crafted high-quality rationales to compile the expert-written pool. We provide a detailed description of the prompt design in Appendix A (Table 4.) Give a new multiple-choice question; we combine the question, the model prediction, and the corresponding extracted facts with the few-shot examples sampled from the expert-pool to formulate the final prompt (see Figure 2.)

3 Evaluation of Rationales

Due to a lack of suitable automated methods for evaluating the rationale quality (Clinciu et al., 2021; Kayser et al., 2021) and credibility, we conducted three studies to address the following questions:

RQ1. How effective are the LLM-generated rationales in communicating KIT model decisions compared to crowdsourced rationales? (§ 4)

RQ2. To what degree do the fine-grained rationale characteristics influence its effectiveness and how generalizable are these observations? (§ 5)

RQ3. How does faithful rationalization of model predictions impact humans' trust in the LLM-generated rationales? (§ 6)

Datasets and Prompts. We select QAGNN (Yasunaga et al., 2021) as the KIT model due to its well-documented code repository and availability of pre-trained model weights. We consider two datasets of multiple-choice QA tasks related to commonsense knowledge, CSQA (Talmor et al., 2019), and elementary-level science, OBQA (Mihaylov et al., 2018). Following the existing KIT models, we use ConceptNet (Speer et al., 2017) as our external knowledge source. For both datasets, we report results on a fixed, randomly-sampled 250-instance test set. We sample these instances from the test set prepared for these datasets (Feng et al., 2020). We employed GPT-3.5 `text-davinci-003` ($temperature = 0$) as the LLM rationalizer. We randomly selected 40 instances from each of the CSQA and OBQA datasets — different from the 250 test instances — to be included in the expert-written example pool. See Appendix A for details.

Faithful Rationalization Studies. We conducted two crowdsourced studies aimed at addressing *RQ1* and *RQ2*. For both studies, we only consider rationalization of correct KIT model predictions, *i.e.*, faithful rationalization. The approach is similar to prior work (Aggarwal et al., 2021; Wiegreffe et al., 2022; Marasovic et al., 2022; Kayser et al., 2021) that also removed the confounder, *i.e.*, rationalization of incorrect model prediction, by only considering rationales for correctly predicted instances. We used Amazon Mechanical Turk for crowdsourcing evaluation. For HITs in both studies, we asked targeted questions to obtain coarse- and fine-grained feedback on the rationales of a KIT model decision. We detail these evaluation metrics in the respective sections discussing the studies. Due to the subjectivity of some of the instances of the CSQA dataset, following Wiegreffe (Wiegreffe et al., 2022), we instruct workers for both the studies to consider the KIT model prediction to be correct even if they disagree with it. We undertook several quality control measures from vetting and recruitment of crowdworkers to accounting for order effect of tasks and individual annotator bias. Besides detailing these measures, we include the study interface design and additional statistical information in Appendices B and E.

Credible Rationalization Study. To address *RQ3*, inspired by existing work on trust in explainable AI (Hoffman et al., 2018; Stites et al., 2021; Smith-Renner et al., 2020), we conducted a confirmatory study in the context of explainable NLP (*i.e.*, LLM-generated rationalization) to explore credibility of rationales on aspects such as agreement, confidence, reliability, and user satisfaction, among others. In this study, we consider rationales generated on both correct and incorrect KIT model predictions. The study was conducted via a Slack campaign within Company X, an industrial research lab, with NLP, data management, and machine learning as the primary research areas.

4 LLMs vs Humans as Rationalizers

We first compare LLM-generated rationales of the CSQA (Talmor et al., 2019) tasks with corresponding crowdsourced rationales from ECQA dataset (Aggarwal et al., 2021). The ECQA rationales are similar in construct to our setting containing corroboration and refutation of CSQA tasks. We exclude CoS-E (Rajani et al., 2019), another crowdsourced free-text rationales dataset, as those rationales are not refutation complete. Moreover, ECQA rationales are reported to be overall better than CoS-E in rationalizing KIT decisions (Aggarwal et al., 2021; Sun et al., 2022). We explain the dataset selection criteria in further detail in Appendix A. Following are the key study takeaways:

- knowledge-guided rationales are preferable (67.2%) to crowdworkers compared to crowd-written rationales, while showcasing substantial increase in preference (45.7%) than prior work (Wiegreffe et al., 2022).
- fine-grained aspects of a rationale such as supportiveness, sufficiency, and convincingness weakly predict such preferences.

4.1 Study Setting

In each of the 250 HITs (three different crowdworkers per HIT), a crowd-worker was presented with a question with choices, the corresponding prediction of the KIT model, and two rationales: LLM-generated (from our pipeline) and crowdworker-written. We then ask them to make a preferential selection among the two rationales (see interface details in Appendix E.1.) We find low-to-moderate annotator agreement – Krippendorff's $\alpha = 0.13$ (Krippendorff, 2011) — for this study, indicating the subjective nature of the task. Related work (Wiegreffe et al., 2022) reported similar agreement statistics ($\alpha \in [0.05, 0.20]$) on comparison between LLM-generated and ECQA rationales.

Fine-grained comparison. Besides head-to-head comparison, we ask several 7-point Likert scale questions — adapted from prior work (Aggarwal et al., 2021; Wiegreffe et al., 2022) — targeted at comparing fine-grained aspects of both rationales. These aspects include: *sufficiency* in justifying the model's choice; *conciseness* (*i.e.*, degree of redundancy); *understandability*; *factuality* (*i.e.*, factual correctnes); *supportiveness* (*i.e.*, the degree to which the model prediction is supported); *refutation convincingness* (*i.e.*, the degree to which the unselected choices are convincingly refuted); *insightfulness* (*i.e.*, how much new information is captured.) New information can be new facts or reasoning not stated in the question and answer choices and potentially grounded on the knowledge evidence. We report the agreement statistics on individual aspects in the Appendix B.2.

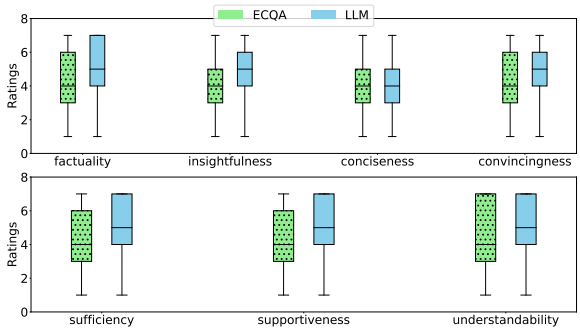


Figure 3: Distribution of fine-grained metrics between crowdworker (ECQA) and LLM-generated rationales — LLM-generated rationales were preferred over ECQA on majority of the metrics except conciseness.

4.2 Higher Preference of LLM Generations

Surprisingly, LLM-generated rationales were more frequently preferred (67.2% times) over crowdworker-written rationales (37.8% times.) The result showcases an improvement over previous work on generating corroboration only (no refutation) rationales (Wiegreffe et al., 2022) — 45.7% preference to LLM generations. The crowdworker-written ECQA rationales potentially outperformed those corroboration rationales on dimensions such as refutation convincingness, sufficiency, and supportiveness. Moreover, our pipeline enabled knowledge-guided rationale generation, whereas prior LLM-generated rationales (Wiegreffe et al., 2022) lacked such grounding and were abstractive. However, the LLMs in both studies differed, with our study employing a newer version (GPT-3.5) than the GPT-3 model used in their work. While some of the performance gain can be attributed to such model upgrades, we demonstrate via fine-grained analysis how aspects of our rationale construct are correlated with crowd-worker preference.

4.3 Fine-grained Comparison

As shown in Figure 3, overall, crowd workers exhibited more preference for LLM-generated rationales over crowdworker-written ones on aspects such as insightfulness (*i.e.*, new information), refutation convincingness, and sufficiency. In fact,

up to 80.4% of the LLM-generated rationales presented to the crowdworkers contained at least one statement grounded on external knowledge, thereby contributing to insightfulness (see Appendix D.1 for details.) Moreover, the refutation argument anchored on the topic of the question enabled a more convincing refutation. Therefore, the resulting LLM-generated rationales were sufficient to justify the model's choice for the QA task.

The preference for LLM-generated and crowdworker-written rationales was comparable for other aspects such as factuality, supportiveness, and understandability. However, crowd-workers rated LLM-generated rationales as more redundant which is unsurprising, given the tendency of the LLMs to generate verbose text.

Metrics	LLM-generated Preferred	Crowdworker-written Preferred
Factuality	0.29	0.04
Insightfulness	0.21	0.12
Conciseness	0.08	0.02
Convincingness	0.29	0.17
Sufficiency	0.28	0.14
Supportiveness	0.27	0.03
Understandability	0.27	0.01

Table 1: Spearman correlation between crowdworker preference of rationales — *weak* correlations are observed with $p < 0.01$ (**strong statistical significance**.)

Correlation to rationale preference. To understand what factors are important for the *preference* judgment, we compute Spearman correlation (Spearman, 1987) between the binary preference of both rationale types — *i.e.*, LLM-generated and crowdworker-written — and the fine-grained aspects (see Table 1.) The conciseness aspect lacked any correlation with either rationale type. Surprisingly, crowd-workers' preference for crowdworker-written rationales lacked any correlation with several other aspects, such as factuality, supportiveness, and understandability, while showcasing a very weak correlation with the rest. However, these fine-grained aspects exhibited a comparatively stronger positive correlation with the LLM-generated rationales. Further analysis showcases that even when crowd-workers preferred ECQA rationales in a head-to-head comparison, LLM-generated and crowdworker-written rationales exhibited almost similar ratings in the majority of the fine-grained aspects (see Appendix D.2.) **Overall**, *the results indicate that human preference for LLM-generated rationale can be captured by factoring in different fine-grained aspects, which can inform the design of automated mechanisms for estimating the suitability of a rationale for end-users.*

5 Acceptability of LLM Rationalization

While pairwise evaluations of preferences provided perspective on the relative quality of the rationales, we conducted another study to independently measure the acceptability of the LLM-generated rationales and collect absolute crowd-worker judgments across several aspects related to rationale quality. We evaluated rationales for both CSQA and OBQA dataset task to understand how generalizable these observations are. The key takeaways from the study are as follows:

- the overall acceptability of the rationales remained high similar to the comparative study.
- however, task and domain variation impacted the quality of the generated rationales.

5.1 Study Setting

In each of the 250 HITs per dataset (three different judges per HIT), a crowd-worker was presented with a question with choices, the corresponding prediction of the KIT model, and an LLM-generated rationale. Besides asking 7-point Likert scale questions on fine-grained aspects of a rationale — similar to the first study in Section 4 — we include two additional surface-level aspects: *readability*, *i.e.*, the clarity of the provided justifications and *grammaticality*, adherence to grammatical rules. Finally, we ask for an overall judgment on quality, *i.e.*, the overall acceptability of a rationale (see interface details in Appendix E.1.) We again find low-to-moderate agreement – Krippendorff's $\alpha = 0.12$ for CSQA and 0.15 for OBQA dataset. Related work (Wiegreffe et al., 2022) reported slightly better agreement statistics ($\alpha = 0.28$) on the CSQA dataset (see Appendix B.2 for details.)

5.2 Favorability Towards LLM generations

On the overall acceptability metric, the LLM-generated rationales received a notably positive rating from the participants for both CSQA ($\mu = 5.83, \sigma = 1.27$) and OBQA ($\mu = 5.89, \sigma = 1.50$). These independent observations reaffirm earlier takeaways (§ 4) and underscore that the LLM-generated rationales of KIT models were viewed favorably by crowd-workers.

Fine-grained observations. As shown in Figure 4, for the newly introduced surface-level metric, readability, and grammaticality, the LLM-generated rationales received higher ratings in keeping with the previous work. In fact, for both datasets, for all of the richer aspects except *insightfulness* and

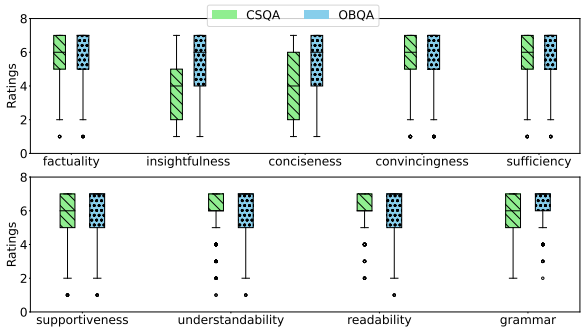


Figure 4: Crowdworkers' ratings showed similar distrbution for all metrics except insightfulness and concisenes. These metrics were rated lower for the more subjective CSQA dataset compared to the objective and scientific OBQA dataset.

conciseness, the ratings received were similar, *i.e.*, more positively rated. While the insightfulness metric was rated positively for OBQA, the rating was neutral to slightly negative for CSQA. Surprisingly, conciseness (*i.e.*, less redundancy) was rated positively for OBQA, whereas CSQA rationales were deemed more redundant, similar to the previous study. A plausible explanation for this discrepancy is the inherent subjectivity in CSQA (Wiegreffe et al., 2022), which can result in varying expectations regarding the information provided in the rationales. In contrast, the OBQA dataset is grounded in objective scientific facts, eliminating such subjectivity and leading to more consistent expectations among crowd-workers.

Metrics	Correlation CSQA	Correlation OBQA
Factuality	0.65	0.73
Insightfulness	0.38	0.67
Conciseness	0.09	0.6
Convincingness	0.70	0.80
Sufficiency	0.76	0.80
Supportiveness	0.54	0.76
Understandability	0.63	0.71
Readability	0.5	0.74
Grammar	0.33	0.62

Table 2: Spearman correlation between acceptability and the fine-grained aspects of a rationale — moderate to fairly strong correlation were observed with **strong statistical significance** ($p < 0.01$).

Correlation to overall acceptability. To understand what factors are important for the overall *acceptability* judgement, we compute Spearman correlation (Spearman, 1987) between *acceptability* and the fine-grained aspects (see Table 2.) For both the datasets, all aspects except *conciseness* show similar patterns — moderate to fairly strong positive correlation with acceptability. However, the rationales for the CSQA dataset (more subjective) exhibited a weaker correlation than the OBQA dataset rationales (more objective) in several aspects, such as conciseness, insightfulness, readability, and grammaticality. **Overall**, *the results indicate that human preference for rationale is more nuanced and can only be holistically captured by considering different fine-grained aspects. However, the quality of the generated rationale may vary depending on the task and domain and, consequently, impact human-preference judgment. Therefore, there is room for improvement in making generated rationales invariant to task and domain.*

6 Towards Credible Rationalization

In the earlier studies, similar to existing work (Aggarwal et al., 2021; Wiegreffe et al., 2022; Marasovic et al., 2022; Kayser et al., 2021), we evaluate LLM-generated rationales for cases where model prediction matches the ground truth. We now investigate the implications of rationalization without accounting for model errors, *i.e.*, faithful rationalization, and potential intervention strategies. Following are the key highlights of the study:

- rationalizing incorrect predictions drastically reduces human's trust in the LLM rationalizer.
- even lightweight guardrails can help intervene more than half of the incorrect predictions.

6.1 Trustworthiness of Generated Rationales

The reported accuracy of KIT models widely vary — 64%-89.4% for CSQA[2] and 60.4%-89.6% for OBQA [3]. The reported human accuracy for the CSQA and OBQA datasets are 88.9% and 91.7%, respectively. Even as humans rationalize, the credibility of the rationalizer may diminish if they attempt to justify any incorrect decisions. Existing work on trust in explainable AI (XAI) literature (Hoff and Bashir, 2015; Schaefer et al., 2016; Stites et al., 2021; Smith-Renner et al., 2020) demonstrates that end-users' trust in a system degrades when encountering errors they can easily recognize due to familiarity and prior experience in a domain. Since the knowledge source for the CSQA and OBQA datasets is ConceptNet (Speer et al., 2017), a commonsense knowledge graph, humans are expected to have higher confidence about their knowledge in the domain. However, existing explainable NLP literature lack studies that investigate the relationship between model accuracy and

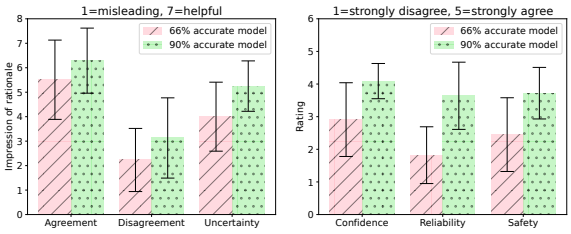
[2] https://www.tau-nlp.sites.tau.ac.il/
[3] https://leaderboard.allenai.org/open_book_qa/

humans' degree of trust in the context of free-text rationales. Therefore, we replicate the study design of exploring trust in explanations for classification models (Stites et al., 2021) to confirm whether the observations hold for knowledge-intensive QA tasks in the commonsense domain.

Study design. We conducted a between-subject study involving 22 participants (15 male and 7 female) exploring two conditions: 66% (11 participants) and 90% (11 participants) model accuracy. The accuracy conditions reflect the two extremities of existing knowledge-intensive task models (Yasunaga et al., 2021; Feng et al., 2020; Yasunaga et al., 2022). The study consisted of three phases: an introduction to the study, a quiz phase, and a follow-up survey. In the quiz phase, the participants answered 15 QA tasks. The 15 tasks were randomly selected from the CSQA (8 QAs) and OBQA (7 QAs) datasets. Depending on the study conditions, for $X\%$ of those N questions, where $X \in \{66, 90\}$, the KIT model made accurate predictions, and the rest of the predictions were inaccurate. The KIT model prediction and LLM-generated rationale of a QA task were revealed *after* a participant submitted their response to avoid bias. Then, the participants were asked whether they agreed with the model prediction and had to rate their impression of the rationale on a scale of 1 to 7 (1 = actively misleading and 7 = helpful.) After the quiz phase, the participants completed a survey adapted from the Trust Scale recommended for XAI (Hoffman et al., 2018). The survey contained questions that asked participants to rate several aspects related to the quiz phase tasks, such as the agreement with rationales and the participants' trust and reliance on the LLM-generated rationale. All of these required participants to work slowly enough to be able to read all the items, thereby making the studies long-running and rather unsuitable for crowd platforms according to existing work (Douglas et al., 2023). Therefore, we opted for internal recruitment as an additional quality control mechanism, inviting participants internally via a Slack campaign at Company X. None of the participants are authors of the paper (see Appendix C.)

6.2 Confirmatory Study Results

The agreement statistics of the participants reflect both the study conditions — 67.27% and 86.07% for lower and higher accuracy models, respectively. Figure 5a summarizes the participants' impression



(a) Impact on user perception (b) XAI Trust Scale feedback

Figure 5: (a) Irrespective of agreement or disagreement with the KIT model prediction, participants indicated a more negative impression about the rationalization of the lower confidence model prediction. (b) Participant feedback on trust scale indicates lower confidence for lower accuracy model rationalization.

of a rationale immediately after viewing the model prediction. When the participants disagreed with the model prediction, they exhibited a stronger negative impression about the rationales for the 66% accuracy condition compared to the 90% accuracy condition. Even when participants agreed with the model prediction, their impression of the rationales remained more negative. Our intuition is that the higher disagreement with the model coupled with observing the faithful rationalization of the incorrect prediction negatively impacted participants' perception of the reliability of the rationales. We confirm these observations by analyzing the results of the follow-up survey (see Figure 5b.) Unsurprisingly, participants for the 66% accuracy condition rated their confidence in the generated rationales and the reliability of the rationalizer significantly lower compared to the 90% accuracy condition. The trends in Figure 5 are observed with strong statistical significance, except for participant feedback on satisfaction with rationale (see Appendix C.2.)

6.3 A Review-then-Rationalize Framework

Motivated by the observations from the preliminary study, we create a two-stage review-then-rationalize (see Figure 6) pipeline to evaluate the impact of intervening incorrect model predictions before rationalization. The pipeline instruments a *reviewer* module that employs another model (GPT-3.5 `text-davinci-003` (`temperature = 0`)) to evaluate the correctness of the knowledge-intensive task model and refrain from rationalizing potentially incorrect decisions.

We opted for LLMs as reviewers due to their reported proficiency in natural language understanding. Depending on the task and data domain, the suitability of the reviewer model may vary. Given

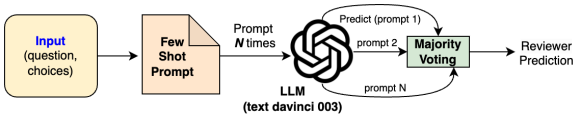


Figure 6: Self-consistency-based Reviewer—intervene for any disagreement with the KIT model prediction.

the complexity of knowledge-intensive tasks, we employ a self-consistency-based decoding strategy (Wang et al., 2022) as opposed to greedy-decoding to ensure robustness. More specifically, we independently pose the same QA task N (=5) times to the reviewer and select the final response via majority voting. The reviewer then compares the model's prediction with its prediction and activates the rationalizer only when both models agree. A cookie-cutter rationale or no rationale may be communicated to the end-user in a disagreement.

Dataset	Prediction Errors (Test Set)	Errors Intervened	
		Greedy Decoding	Self-consistency
CSQA	321	166 (51.71%)	187 (**58.26%**)
OBQA	155	102 (65.81%)	110 (**70.97%**)

Table 3: The review-then-rationalize pipeline helps intervene incorrect predictions of a knowledge-intensive task model. The self-consistency-based reviewer outperforms the greedy decoding-based reviewer.

As shown in Table 3, for knowledge-intensive tasks such as CSQA and OBQA, the proposed pipeline helps intervene up to 58% and 71% of the incorrect predictions. Unsurprisingly, the self-consistency-based reviewer outperforms the greedy decoding-based reviewer. **Overall**, *the results draw attention to the importance of responsibly communicating LLM-generated rationales to humans and consequently, instrumenting guardrails as an effective intervention strategy.*

7 Related Work

Free-text Rationale Generation. Existing works highlight the effectiveness of free-text rationales in justifying a model's decision to humans in vision (Hendricks et al., 2016; Park et al., 2018) and text domains (Camburu et al., 2018; Ehsan et al., 2018; Narang et al., 2020). Due to cost and generalizability implications of supervised rationale generation, we employ few-shot prompting to elicit rationales from LLMs following existing work (Wiegreffe et al., 2022; Marasović et al., 2021). Both these approaches generate abstractive, corroborative, and faithful rationales. In contrast, we explore the generation of knowledge-guided, corroborative and refutation complete, and credible rationales.

Guided text generation. Developing approaches to avoid hallucinations and factual inaccuracies in LLM-generated text is a new area of research. Retrieval augmented generation (RAG) infuses external knowledge (Peng et al., 2023; Lazaridou et al., 2022), such as knowledge-bases and web documents, while prompting LLMs to help generate responses. We employ a similar strategy during rationalization by conditioning the LLM generation on the retrieved evidence for a given task.

Credible text generation. Studies in explainable AI literature (Smith-Renner et al., 2020; Hoff and Bashir, 2015; Schaefer et al., 2016; Stites et al., 2021) demonstrate that for low-quality models, providing faithful explanations further degraded user's trust. Unlike existing work on free-text explanation (Wiegreffe et al., 2022; Marasović et al., 2021), we explore how end-users' trust may be impacted by faithful rationalization of varying degrees of incorrect model predictions. ReXC (Majumder et al., 2021) augments rationales — generated in a self-rationalization framework — with background knowledge to improve a model's task performance, such as natural language inference and visual commonsense reasoning. To rectify incorrect LLM responses, identified via a self-consistency-based intervention approach, the Verify-then-Edit framework (Zhao et al., 2023) leverages external knowledge to repair reasoning chains of the corresponding chain-of-thought prompts. FARM (Mei et al., 2023) utilizes trustworthy external sources within a predict-then-generate framework that aims to intervene in harmful content generation using LLMs. To credibly rationalize KIT model predictions, we explore a review-then-rationalize framework where a self-consistency-based reviewing approach identifies potential prediction inaccuracies and ensures credible rationale generation.

8 Conclusion

We evaluate LLMs' capacity to generate effective rationales for knowledge-intensive tasks in a few-shot knowledge-guided setting. We additionally investigate the implications of employing LLMs as rationalizers of an imperfect model and highlight the negative impact on users' trust. Observations from our studies highlight room for improvement in aspects such as task and domain invariant rationalization and robust intervention strategies for real-world usage.

9 Limitations

Scrutinizing LLM-generated rationales. While external knowledge-guided generation offers promise (Peng et al., 2023; Mallen et al., 2023), LLM-generated rationales may still suffer from hallucinations. Our experiments highlight that the LLM-generated rationale is not entirely grounded on retrieved knowledge. Even though crowdworkers positively rated the factuality and insightfulness of the generated rationales, additional scrutiny is required before deploying such rationalizers in mission-critical tasks. To this end, the review-then-rationalize framework may be expanded to further scrutinize the rationales by adopting recent work on an LLM's factual knowledge measurement (Pezeshkpour, 2023; Dong et al., 2023) and hallucination identification (Manakul et al., 2023; Elaraby et al., 2023; Mündler et al., 2023) and reduction (Zhao et al., 2023; Mei et al., 2023), and explainable evaluation (Xu et al., 2023).

Fairwashing vs. credible rationalization. The accuracy of our self-consistency-based reviewer can be further improved to intervene in a higher proportion of incorrect KIT model predictions. However, critiques of XAI tools have raised concerns about *fairwashing*, *i.e.*, misleading users into trusting biased or incorrect models (Alikhademi et al., 2021). For example, simply averting potential faithful yet incorrect rationalization, identified by the reviewer, may increase end-users' trust due to an illusion of a highly performant rationalizer (Aïvodji et al., 2019). Such fairwashing may have catastrophic consequences if employed in real-world applications such as in the medical domain, hiring platforms, and credit agencies. Recent work (Alikhademi et al., 2021) proposes a framework for evaluating XAI tools with respect to their capabilities for detecting and addressing issues of bias and fairness as well as their capacity to communicate these results to their users clearly. Therefore, future implementations of the credible rationale should adopt similar strategies to safeguard against such issue. Future work may explore different communication strategies during prediction errors, such as communicating the disagreement to the experts-in-the-loop, providing rationales with a disclaimer, and employing stronger reviewers to repair the prediction on the fly and then rationalize, among others.

Scaling responsibly. An often overlooked aspect of the recent popularity of LLMs has been *Green AI* (Schwartz et al., 2020). When the ML deployment pipeline is considered as a whole, inference consumes most compute resources, accounting for anything between 70% to 90% (Weng et al., 2022; Wu et al., 2022). Knowledge distillation approaches can be adopted to avoid costly pretraining (Wang et al., 2023). Furthermore, materialization of rationales to avoid repeating rationalizing the same task can be possible approaches to handle such issues.

Crowdsourcing study constraints. As we conducted the crowdsourced study on Amazon Mechanical Turk, our findings may not generalize to other platforms and feedback provided in in-person lab-based studies. Moreover, we observed low agreement among the annotators — similar to prior work (Wiegreffe et al., 2022) — due to the subjectivity of the QA tasks. Future work may explore conducting large-scale studies with better quality control mechanisms (such as hiring private firms with dedicated teams similar to (Aggarwal et al., 2021) and conducting in-house studies with experts. Such a setting also makes to possible to collect additional insights into the thought process of the participants. However, conducting such large-scale studies in an in-person setup introduces time and logistics constraints. To this end, recent LLM-based reference-free approaches (Liu et al., 2023) to scale-up evaluation offers promise. Although, it is unclear whether such evaluation strategies apply to subjective metrics of rationale quality studied in our work. Therefore, future studies may explore how such reference-free judgements align with human judgements similar to (Pezeshkpour, 2023).

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)*, Online. Association for Computational Linguistics.

Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR.

Kiana Alikhademi, Brianna Richardson, Emma Drobina, and Juan E Gilbert. 2021. Can explainable ai explain

unfairness? a framework for evaluating explainable ai. *arXiv preprint arXiv:2106.07483*.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

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

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. 2023. Statistical knowledge assessment for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Benjamin D Douglas, Patrick J Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between mturk, prolific, cloudresearch, qualtrics, and sona. *Plos one*, 18(3):e0279720.

Upol Ehsan, Brent Harrison, Larry Chan, and Mark O Riedl. 2018. Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, Laura Freeman, and Feras A Batarseh. 2023. Rationalization for explainable nlp: A survey. *arXiv preprint arXiv:2301.08912*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 3–19. Springer.

Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434.

Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *arXiv preprint arXiv:2310.11207*.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1244–1254.

John King and Roger Magoulas. 2015. *2015 data science salary survey*. O'Reilly Media, Incorporated, Sebastopol, CA, USA.

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

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. 2021. Knowledge-grounded self-rationalization via extractive and natural language explanations. *arXiv preprint arXiv:2106.13876*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, Seattle, United States. Association for Computational Linguistics.

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

Alex Mei, Sharon Levy, and William Yang Wang. 2023. Foveate, attribute, and rationalize: Towards physically safe and trustworthy ai. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11021–11036.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

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*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models. *arXiv preprint arXiv:2306.06264*.

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.

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.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2021. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3):377–400.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems*, pages 1–13.

Charles Spearman. 1987. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Mallory C Stites, Megan Nyre-Yu, Blake Moss, Charles Smutz, and Michael R Smith. 2021. Sage advice? the impacts of explanations for machine learning models on human decision-making in spam detection. In *International Conference on Human-Computer Interaction*, pages 269–284. Springer.

Jiao Sun, Swabha Swayamdipta, Jonathan May, and Xuezhe Ma. 2022. Investigating the benefits of free-form rationales. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5867–5882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023. Scott: Self-consistent chain-of-thought distillation. *arXiv preprint arXiv:2305.01879*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Samuel F Way, Daniel B Larremore, and Aaron Clauset. 2016. Gender, productivity, and prestige in computer science faculty hiring networks. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1169–1179, New York, NY, USA. ACM.

Qizhen Weng, Wencong Xiao, Yinghao Yu, Wei Wang, Cheng Wang, Jian He, Yong Li, Liping Zhang, Wei Lin, and Yu Ding. 2022. {MLaaS} in the wild: Workload analysis and scheduling in {Large-Scale} heterogeneous {GPU} clusters. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22)*, pages 945–960.

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 Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. *arXiv preprint arXiv:2102.12060*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2023. Less is more for long document summary evaluation by llms. *arXiv preprint arXiv:2309.07382*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback. *arXiv preprint arXiv:2305.14282*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*.

A Prompts and Rationales

In this section, we provide additional details regarding the prompts corresponding to the faithful and credible rationalization workflows.

A.1 Faithful Rationalization

Figure 7 outlines the few-shot prompt structure. Table 4 elaborates on the prompt design shown in (Figure 2 and Figure 7). Each example in the few shot prompt includes the question and answer choices, the KIT model selected answer, the knowledge facts extracted from ConceptNet for each choice, and the expert-written question topic and rationale that act as input to GPT-3.5 `text davinci 003`. While we show only two few-shot examples, in practice, we use five examples per prompt. As explained in the Section 2, due the token limit imposed by the GPT-3.5 API, we can include from 5-8 examples depending on the length of the knowledge facts. Given the prompt, *i.e.*, examples followed by an unseen question and answer choices, KIT model selected answer, and extracted knowledge, the LLM greedily generates the question topic and the rationale for the model prediction.

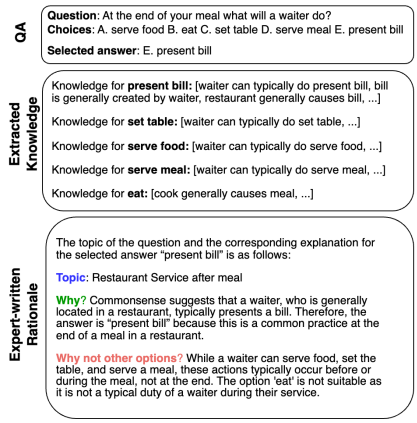


Figure 7: An example in the few-shot prompt: the QA and External Knowledge components are retrieved and the topic and the rationale are expert authored.

To design the initial prompt, we take inspiration from existing work (Wiegreffe et al., 2022; Peng et al., 2023; Lazaridou et al., 2022; Zhao et al., 2023) to experiment with the prompt layout. We experimented with approximately 6 different layouts in the OpenAI playground [4] using 10 training examples on the CSQA and OBQA datasets. In deciding the number of few-shot examples, we considered the maximum context window size of GPT-3.5 `text-davinci-003`, which is 4097 tokens. We observed that depending on the datasets and the length of the factual statements retrieved from ConceptNet, five to eight few-shot examples fit into the token constraints. After finalizing the prompt layout, we developed a pool of 40 expert-written (*i.e.*, authors of these papers) examples. We randomly selected 5 expert-written examples for each test instance to ensure uniformity across datasets and instances. Similar to prior work (Wiegreffe et al., 2022), we focused on developing a general few-shot prompting strategy for generating knowledge-enhanced and refutation complete rationale that could be successful when no additional (large) validation set for parameter tuning is available. We prompt the LLM to generate a topic of the question and a rationale similar to the provided few-shot examples. Therefore, our approach explicitly conditions the rationale generation on the question topic and the knowledge facts. FARM (Mei et al., 2023) employs a similar topic-focused generation for question answering. Given a question, the LLM is initially prompted to generate a question context — augmented with information retrieved from trustworthy sources — to generate a safe response. Such strategies have been shown to be very effective (Radford et al., 2019; Brown et al., 2020; Shin et al., 2020; Schick and Schütze, 2020), even in complex generation tasks (Reif et al., 2021).

Relevance to ECQA rationales. The pipeline for ECQA (Aggarwal et al., 2021) rationale generation and the knowledge-guided LLM rationalization have several similarities. As shown in Table 4, the rationalization pipeline provides conceptnet assertions corresponding to the selected answer and rejected choices as context within the prompt. ECQA crowdsourcing pipeline also prompted crowdworkers to use the positive facts about the selected answer and negative facts about the other choices as guides to craft the eventual free-flow explanation. Therefore, in both cases, the rationalizer pipeline, be it crowdworker or LLMs, were knowledge-guided. However, in case of LLMs the source of knowledge guidance is external, *i.e.*, Concept-

[4] https://platform.openai.com/playground

Question: At the end of your meal what will a waiter do?
Choices: A. serve food B. eat C. set table D. serve meal E. present bill
Selected answer: E. present bill

Knowledge for present bill: [waiter can typically do present bill, bill is generally created by waiter, restaurant generally causes bill, . . .]
Knowledge for set table: [waiter can typically do set table, . . .]
Knowledge for serve food: [waiter can typically do serve food, . . .]
Knowledge for serve meal: [waiter can typically do serve meal, . . .]
Knowledge for eat: [cook generally causes meal, . . .]

The topic of the question and the corresponding explanation for the selected answer "present bill" are as follows:
Topic: Restaurant Service after meal
Why? Commonsense suggests that a waiter, who is generally located in a restaurant, typically presents a bill. Therefore, the answer is "present bill" because this is a common practice at the end of a meal in a restaurant.
Why not other options? While a waiter can serve food, set the table, and serve a meal, these actions typically occur before or during the meal, not at the end. The option 'eat' is not suitable as it is not a typical duty of a waiter during their service.
==========
Question: He waited for his friend at the squash court, but he was worried his friend thought he meant the at the other end of the public what?
Choices: A. country club B. rich person's house C. pool D. park E. fitness center Selected Answer: D. park

Knowledge for park :[squash court is generally located in park, play is generally located in squash court, . . .]
Knowledge for fitness center :[squash court is generally located in fitness center, . . .]
Knowledge for country club :[squash court is generally located in country club, . . .]
Knowledge for pool :[. . .]
Knowledge for rich person's house :[. . .]

The topic of the question and the corresponding explanation for the selected answer "park" are as follows:
topic: Public Spaces and Miscommunication
Why? The answer is park because commonsense suggests that a squash court is generally located in a park. This implies that there could be another squash court at the other end of the park, leading to the friend's confusion.
Why not other options? While a squash court can also be located in a fitness center or country club, these locations are not typically public spaces with multiple ends. A pool or a rich person's house are less likely to have multiple squash courts, making them less likely to be the source of the friend's confusion.
==========
Question: What should the bean bag chair sit on?
Choices: A. house B. den C. family room D. wood E. floor
Selected Answer: E. floor

Knowledge for present floor: [. . .]
Knowledge for house: [. . .]
Knowledge for den: [. . .]
Knowledge for family room: [. . .]
Knowledge for wood: [. . .]

The topic of the question and the corresponding explanation for the selected answer "present bill" are as follows:

Table 4: Example of a prompt with two training examples for CSQA and an unseen question for which the LLM generated a rationale. In practice, we provide five examples.

Net (Speer et al., 2017), whereas for ECQA, the crowdworkers themselves crafted the supporting facts before rationalizing.

A.2 Credible Rationalization

Table 5 showcases the prompt design for the *Reviewer* model within the credible rationalizer pipeline (Figure 6. Each of the five examples in the few shot prompt includes the question and answer choices that act as input to GPT-3.5 `text davinci 003`. In practice, we use five examples per prompt. Given the prompt, *i.e.*, examples followed by an unseen question and answer choices, the LLM greedily generates a response, *i.e.*, predicts an answer from the choices. We repeat the process five times and select a response based on majority voting. We randomly sample five questions from the 40 expert-written rationale pool as few-shot examples.

Question: At the end of your meal what will a waiter do?
Choices: A. serve food B. eat C. set table D. serve meal E. present bill
Selected answer: E. present bill
==========
Question: He waited for his friend at the squash court, but he was worried his friend thought he meant the at the other end of the public what?
Choices: A. country club B. rich person's house C. pool D. park E. fitness center Selected Answer: D. park
==========
Question: What should the bean bag chair sit on?
Choices: A. house B. den C. family room D. wood E. floor
Selected Answer:

Table 5: A prompt with two training examples for CSQA and an unseen question for the Reviewer to answer.

A.3 LLM-generated Rationales

Table 6 a few non-cherry picked examples of LLM-generated rationales. We show examples of rationales for CSQA dataset questions generated by both LLM and humans, *i.e.*, crowdworkers in the ECQA dataset. Since there is no crowdsourced dataset of OBQA rationales, we only show LLM-generated rationales. Note that the LLM greedily

generates a topic of the question and a rationale with corroboration ("Why?") and refutation ("Why Not") components. However, as shown in Figure 2, the these two components are extracted from the generated output to construct the eventual rationales (*i.e.*, the formatting step at the end.)

==========
CSQA Dataset
==========
Question: What should the bean bag chair sit on?
Choices: A. house B. den C. family room D. wood E. floor
Selected Answer: E. floor

LLM-generated rationale: The answer is floor because the commonsense knowledge clearly indicates that a bean bag chair is generally located in a floor.
While a bean bag chair can be placed in a house, den, family room, or on wood, the floor is the most common place for a bean bag chair to be located.

ECQA rationale: Bean bag chair is a seat people sit on which is generally put on a floor. A bean bag chair should sit on a floor and not on anything else from the other options.

==========
OBQA Dataset
==========
Question: Rainbows are always found after what?
Choices: A. A fire B. A tornado C. Rainfall D. Cereal
Selected Answer: C. Rainfall

LLM-generated rationale: The answer is Rainfall because rainbows are always found after rain. This is because the sunlight is refracted by the raindrops in the air, creating the rainbow. A fire, a tornado, and cereal do not have any relation to rainbows.

Table 6: Rationales for CSQA generated by LLM and humans (ECQA) and for OBQA generated by LLM.

B Crowd Study Details

We provide more details regarding both crowdworker studies such as additional statistics related to the crowd study and quality control mechanisms.

B.1 Quality Control and Payment

In order to enforce quality throughout evaluation, we use a hidden built-in Javascript function to compute time per HIT spent and perform attention checks by inserting questions with specific instructions randomly within a HIT. We filter out any annotator who completed the tasks in an unreasonably low time, or failed their attention checks. To mitigate individual annotator bias, we also ensure that each experiment in a study has a substantial number of distinct crowdworkers. See Tables 7 and 8 for details regarding the inter-annotaror agreement for the comparison study. For both studies, we used a pay rate of USD 12.00/hr. We performed periodic check to ensure that the median HIT completion time remains commensurate to approximately the pay rate. Median times reported for the comparative study was 208 seconds (paid at 80 cents each) the acceptability study was 110 seconds (paid at 40 cents each.) To ensure the quality of responses, we require annotators in Australia, New Zealand, United Kingdom, United States, and Canada as a proxy for English competency. We only selected workers with a past approval rate $> 98\%$ and who have completed over 5000 HITs. We documented a worker's HIT submission time and performed attention checks within each HIT to enforce quality control. Note that each crowd worker was presented with detailed instructions about the study interface and performed an example task as a warm-up.

Approach	LLM-generated	ECQA
Factuality	0.07	0.05
Insightfulness	0.15	0.03
Conciseness	-0.04	-0.01
Convincingness	0.09	0.03
Sufficiency	0.08	0.07
Support	0.08	-0.01
Understandability	0.09	0.06
Preference	0.13	0.13

Table 7: Inter annotator agreement (Krippendorff's α) of crowdworkers on the fine-grained aspects of a rationale evaluated in the head-to-head comparison study.

Dataset	CSQA	OBQA
Factual	0.02	0.03
Insightful	-0.06	-0.04
Concise	-0.15	-0.17
Convincing	0.08	0.13
Sufficient	0.07	0.08
Support	-0.012	-0.002
Understandable	0.02	0.04
Readability	-0.05	-0.02
Grammar	-0.15	-0.16
Acceptability	0.12	0.15

Table 8: Inter annotator agreement (Krippendorff's α) of crowdworkers on all the coarse- and fine-grained aspects of a rationale evaluated in the acceptability study.

B.2 Annotator Statistics

We now report the number of distinct crowd annotators and the median and mean number of HITs completed for each experiment. For the head-to-head comparison study, there were 750 HITs in total. There were 29 unique annotators with a median of 10 (mean = 21.86) HITs completed by an annotator. For the acceptability study, there 750 HITs for each of the two datasets CSQA and OBQA. For the CSQA dataset, there were 25 unique annotators with a median of 7 (mean = 28.80) HITs completed by an annotator. For the OBQA dataset, there were 30 unique annotators with a median of 7 (mean = 25.00) HITs completed by an annotator. More detailed breakdowns of inter-annotator agreement for both studies are reported in Tables 7 and 8.

C Credible Rationalization Study

We now provide relevant information complementing the observations obtained in the preliminary study regarding credible rationalization.

C.1 Study Details

Participants. The participants of the preliminary study were all from Company X. However, we still performed attention checks in the preliminary study. The participants were unaware of the hypothesis and evaluation objective of the study. None of the participants are authors of the paper. Out of the 20 participants in the study, 15 were male and 5 were female. The representation of the female participants (25%) compares favorably with recent estimates of 15% women in tenure-track faculty in computing (Way et al., 2016) and 20% women in data science positions worldwide (King and Magoulas, 2015). One-fourth of the participants held a Bachelor degree and the rest completed graduate school or higher. Due to the complexity and longer duration of this study, we wanted to ensure the participation of higher quality participants by such selective recruitment.

Phases. We first collected participants' demographic information and then provided detailed instructions about the subsequent phases: a quiz phase consisting of a collection of tasks and a follow-up survey. The survey is adapted from the Trust Scale recommended for XAI (Hoffman et al., 2018). We opted for a follow-up survey rather than after each task completion following Hoffman et al. (Hoffman et al., 2018) — "the questions are appropriate for scaling after a period of use, rather than immediately after a rationale has been given." Besides questions related to the trust scale, we also asked participants to rate their overall acceptability of the rationales on a scale of 1 to 5. Note that the acceptability rating scale is different from the earlier studies in Section 5 and 4 to conform with the Trust Scale ratings (Hoffman et al., 2018).

C.2 Feedback Statistics

We conducted *Mann-Whitney U test* to measure the statistical significance of the differences between the 66% and 90% accurate model conditions, along various credibility metrics proposed in Section 6. The Mann-Whitney U test is a non-parametric test to measure the significance of difference in distribution of two independent sample, *i.e.*, accuracy conditions in this study.

As shown in Table 9, participant feedback on individual task indicates a higher disagreement with lower confidence model prediction and a more negative impression regarding the corresponding rationale. The differences is significant both cases *i.e.*, when participants either agreed or disagree with the KIT model prediction. Table 10 reports the summary of participant feedback during the post-quiz survey — participants exhibited a more negative impression regarding the corresponding rationale. For all of the aspects except *statisfaction*, the difference in participant feedback between the accuracy conditions were statistically significant.

Table 11 summarizes the observations from the quiz phase, *i.e.*, participant agreement statistics with the model prediction and participants' impression of the corresponding rationale. The agreement statistics (overall = 76.67%) of the participants reflect both the study conditions — 67.27% and 86.07%, respectively. Due to the subjective nature of the tasks, especially in the CSQA dataset, a few participants were unsure whether to agree or disagree with the model predictions, further reflecting the difficulty of the tasks.

D Additional Experiments and Analysis

We now present details of various user study observations, discussed briefly in earlier sections.

D.1 Degree of Knowledge Grounding

While our proposed knowledge-graph-based retrieval augmented LLM-generated rationales were positively rated by crowdworkers, questions remain regarding the effectiveness of such knowledge grounding. To evaluate whether any fragments of the rationales generated using our proposed approach were grounded on the retrieved knowledge facts, we conducted an experiment. We primarily focus on the corroboration component as there is a higher probability of the knowledge graph containing facts about the correct answer choice.

We measure the existence of knowledge-grounding as follows: consider the retrieved knowledge corresponding to the correct choice j for question q_i in dataset D, $\mathcal{G}_{ij}$, and the corroboration component of the corresponding LLM-generated rationale, RC_i. We first measure the BERTScore (Zhang et al., 2019) similarity between a fact $f \in \mathcal{G}_{ij}$, expressed in natural language and a sentence $s \in RC_i$. We then select the fact-sentence pair, (f, s), with the highest BERTScore as a po-

Agreement = yes (†)			Agreement = no (*)			Agreement = unsure		
Accuracy 66%	**Accuracy** 90%	**Stat. Sig.**	**Accuracy** 66%	**Accuracy** 90%	**Stat. Sig.**	**Accuracy** 66%	**Accuracy** 90%	**Stat. Sig.**
$\eta = 6.00$	$\eta = 7.00$		$\eta = 2.00$	$\eta = 3.00$		$\eta = 4.00$	$\eta = 5.00$	
$\mu = 5.89$	$\mu = 6.29$	$\mathbf{p < 0.01}$	$\mu = 2.23$	$\mu = 3.13$	$\mathbf{p < 0.05}$	$\mu = 4.00$	$\mu = 5.25$	$p > 0.05$
$\sigma = 1.62$	$\sigma = 1.33$		$\sigma = 1.29$	$\sigma = 1.64$		$\sigma = 1.41$	$\sigma = 1.03$	

Table 9: Participant feedback on individual task indicates a more negative impression — rated on a scale between 1 (misleading) to 7 (helpful) — regarding the corresponding rationale. (†) indicates statistical significance with $pa < 0.01$ and (*) indicates statistical significance with $p < 0.05$.

Metric	**Confidence** (†)		**Reliability** (†)		**Safety** (†)		**Satisfaction**		**Acceptability** (†)	
Accuracy	66%	90%	66%	90%	66%	90%	66%	90%	66%	90%
Median	3.00	4.00	2.00	4.00	3.00	4.00	3.00	5.00	3.00	4.0
Mean	2.91	4.09	1.82	3.64	2.45	3.72	3.45	4.55	3.09	4.27
Std. Dev.	1.14	0.54	0.87	1.03	1.13	0.79	1.44	0.52	1.04	0.65

Table 10: Participant feedback on individual task indicates a more negative impression regarding the corresponding rationale. (†) indicates statistical significance with $p < 0.01$.

Metric	**Agreement** %		
	Overall	**Accuracy** 66%	**Accuracy** 90%
Agreement = yes	76.67%	67.27%	86.07%
Agreement = no	20.30%	31.52%	9.09%
Agreement = unsure	3.03%	1.21%	4.85%

Table 11: Participant feedback on individual task indicates a higher disagreement with lower confidence model prediction.

Dataset	**Pairwise Max BERTScore**	**Percentage of Entailment**
CSQA	$\mu = 0.5823, \sigma = 0.0650$	80.4%
OBQA	$\mu = 0.5173, \sigma = 0.0803$	38%

Table 12: Degree of knowledge grounding observed in the LLM-generated rationales.

tential candidate for evaluating whether the fact f entails the sentence s within the rationale. Such entailment is an indicator of whether a fragment of a rationale being grounded on retrieved knowledge facts. Similar approach has been adopted in existing work (Wu et al., 2023) to extract candidate sentences from long documents and evaluate the degree to which the corresponding summary is grounded on the source document. Following their approach, we employ NLI models (Reimers and Gurevych, 2019), *i.e.*, DeBERTa-base model fine-tuned on SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018), to evaluate entailment. For the BERTScore, we used DeBERTa-Large model (He et al., 2020) fine-tuned on MNLI.

We measure the knowledge-grounding statistics of the CSQA and OBQA dataset rationales evaluated in the acceptability crowd study in Section 5. As shown in Table 12, on average, at least one fact-sentence pair achieved BERTScore of 0.5823 and 0.5173 for CSQA and OBQA datasets, respectively. While a higher percentage of those pairs were classified as entailment (80.4%) for CSQA, the entailment statistics was a bit lower for OBQA. On reflection, the lower value seems reasonable since we used ConceptNet, a commonsense knowlege graph, as the external source for OBQA, a dataset on elementary science question answering.

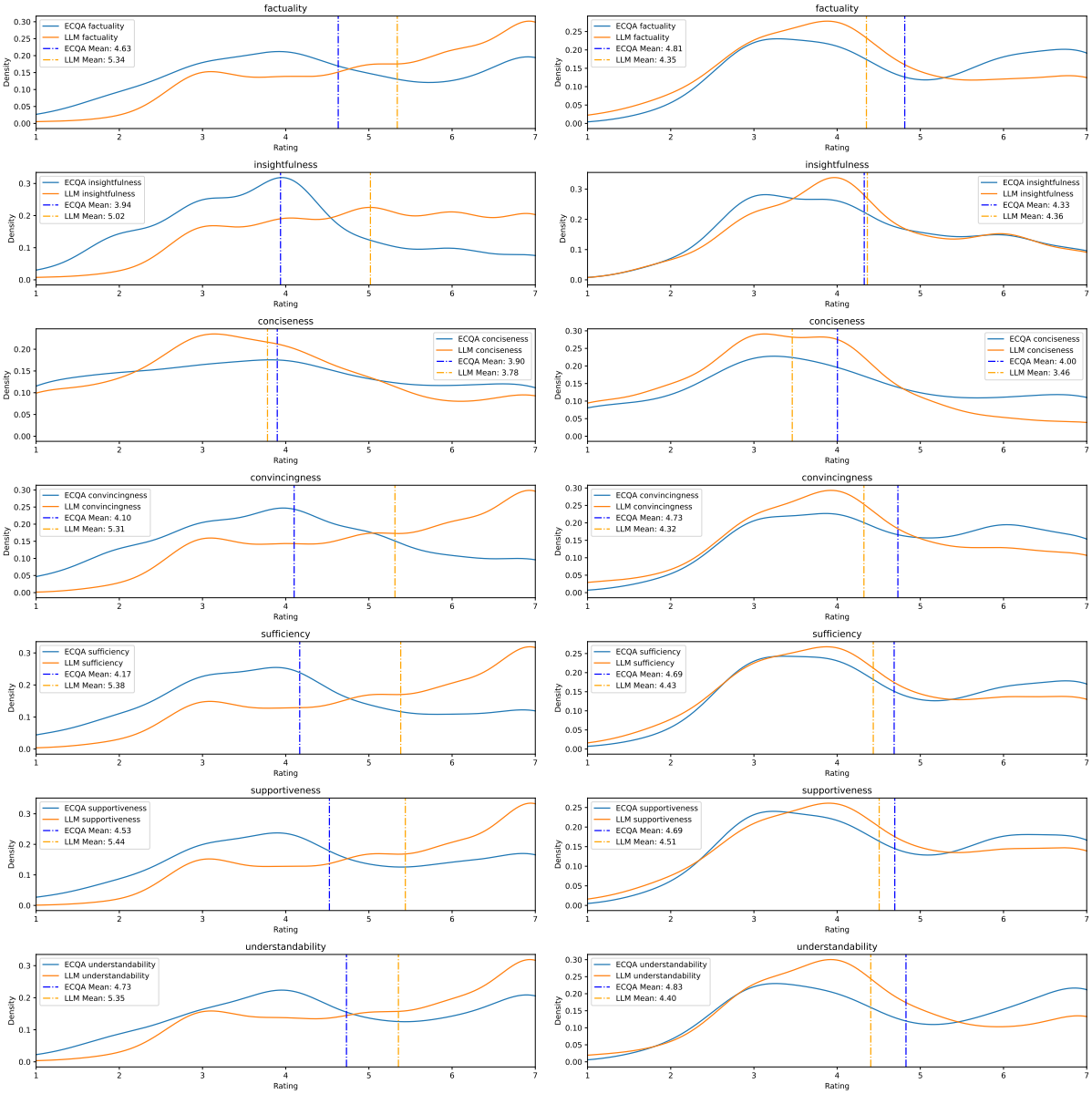
The initial observations highlight the promise of knowledge-guided rationalization in ensuring factuality of LLM-generated rationales. However, more in-depth analysis with a stronger metric that takes into account multiple fact-sentence pair candidates across corroboration and refutation components is required to reliably capture the degree of knowledge. Such fine-grained analysis is beyond the scope of our study and can be explored in future.

D.2 A Deeper Dive into LLM vs ECQA

To better understand, we further analyze the crowd worker feedback based on their preference of rationales. Cases where workers preferred LLM-generated rationales over humans (*i.e.*, the 61.8% cases) — LLM-generated rationales were rated substantially higher than human-written rationales, except conciseness (see Figure 8.) Even the conciseness rating for both types of rationales was almost the same, with human-written rationales faring slightly better. On the other hand, for cases where workers preferred human-written rationales over LLMs (*i.e.*, the 38.2% cases) — surprisingly, apart from conciseness, human-written rationales were rated significantly higher only on two aspects: factuality and convincingness. For the rest of the aspects, the differences between ratings of both rationale types were marginal.

E Study Interfaces

In this section, we provide screenshots of the important aspects of the three studies.



(a) LLM preferred (b) ECQA preferred

Figure 8: (a) LLM-generated rationales preferred over human-written (ECQA) rationales. LLM-generated rationales were rated substantially higher than human-written rationales, with the exception of conciseness. (b) ECQA rationales preferred over LLM-generated rationales. Surprisingly, human-written rationales were rated significantly higher only on three aspects: conciseness, factuality and convincingness.

E.1 Faithful Rationalization Interface Details

Both studies were conducted in the Amazon Mehcanical Turk. We mentioned the worker inclusion criteria in Section 3. Each study was launched in separate batches and were not conducted simultaneously. Due to the complexity of HITs in each of the studies, we designed the study interfaces from scratch using HTML and JavaScript. These interfaces were uploaded in the platform as a new project to launch the corresponding study.

Figure 9 shows a screen shot of the HIT interface of the first study — head-to-head comparison between LLM-generated and ECQA (crowdworker-written) rationales. The HIT contains a question and the choices, a selected answer, and two rationales, order of which the are determined at random on-the-fly. Figure 10 shows a screen shot of the HIT interface of the acceptability crowd study with a question and the choices, a selected answer, and an LLM-generated rationales. For both the studies, the workers were asked several rating questions de-

signed to collect feedback on both coarse-grained and fine-grained aspects of a rationale outlined in Section 4 and Section 5. Workers were asked to rate the rationale(s) using a sliding scale ($[1, 7]$).

We opted for Likert scale-based rating rather than choice questions to get a more fine-grained feedback. Given a choice questions, each choice may not exactly capture the participants interpretation of how much a rationale observed the property being evaluated. For example, as shown in Figure 9, we ask the crowdworker to "rate how understandable each rationale is ". To assist the participants, we suggest how to use the scale — provide interpretation of three points in the scale, *i.e.*, 1 = Not understandable, 4 = Somewhat understandable, and 7 = Completely understandable.

Additional quality control measures. Note that some instances in CSQA have multiple correct or very similar answer choices, due to noise in the dataset and the fact that the wrong answer choices were deliberately collected to make the task challenging. To remove this possible confounder, following related work (Wiegreffe et al., 2022), in both the crowd studies we instruct crowdworkers to treat the selected answer as correct even if they disagree with it, and then rate the fine-grained aspects of the rationales. To minimize bias, we randomized the order in which rationales were displayed side-by-side across workers of each HIT. We also randomized randomized the order of the rating questions on the fine-grained aspects presented across workers of each HIT. Three different workers completed each HIT. The workers who participated in the comparative study were excluded from the acceptability study. Furthermore, for the acceptability study, we launched the OBQA dataset HITs after the conclusion of the CSQA HITs and excluded workers participating in the CSQA HITs.

E.2 Credible Rationalization Interface Details

As shown in Figure 11, participants are first asked to answer a multiple choice question sampled randomly from the CSQA and OBQA datasets. We ensure the accurate distribution of questions with correct and incorrect KIT model prediction for each study condition by grouping questions in each dataset by prediction accuracy. Once the participant selects an answer, they are shown the KIT model prediction and the LLM-generated rationale (Figure 12). At this point, the QA component is disabled so the the participants cannot change their options. Finally, participants are provided two follow up questions to collect immediate feedback regarding the task (Figure 13). Finally, participant conclude the study by completing a survey with questions adapted from the XAI trust scale (Hoff and Bashir, 2015) (see Figrue 14.)

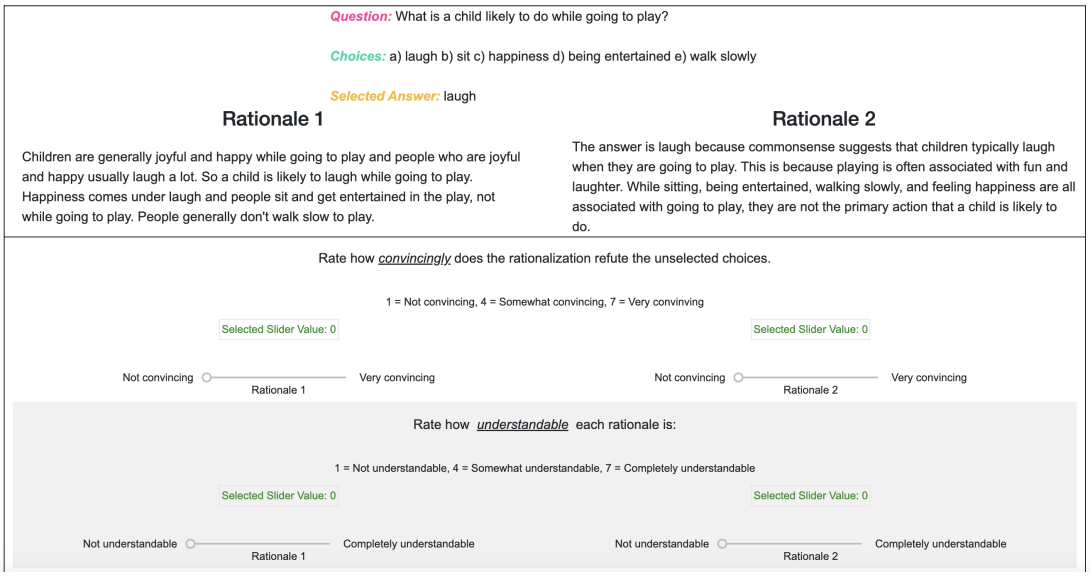


Figure 9: A partial screenshot of the head-to-head comparison interface.

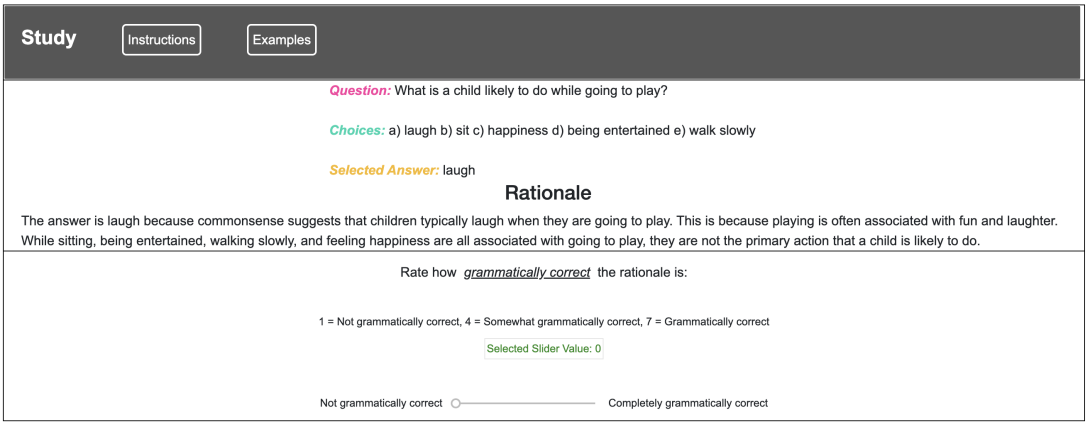


Figure 10: A partial screenshot of the acceptability task interface.

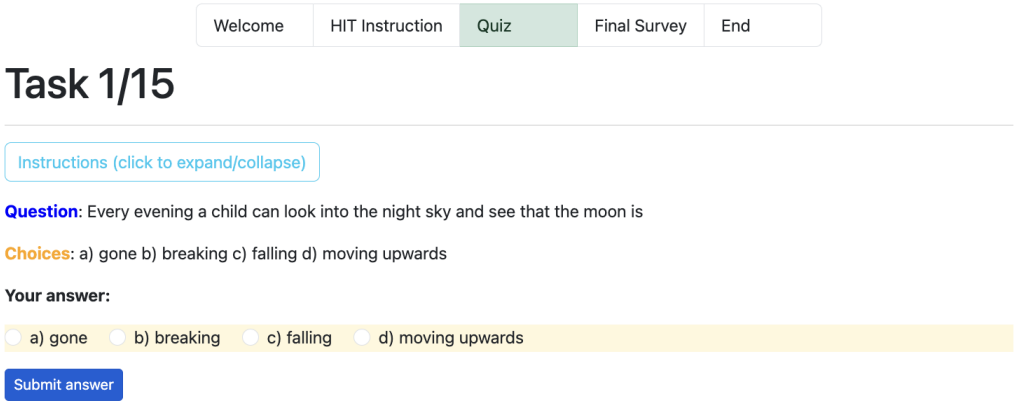


Figure 11: For each task, participants are first asked to answer a multiple choice question.

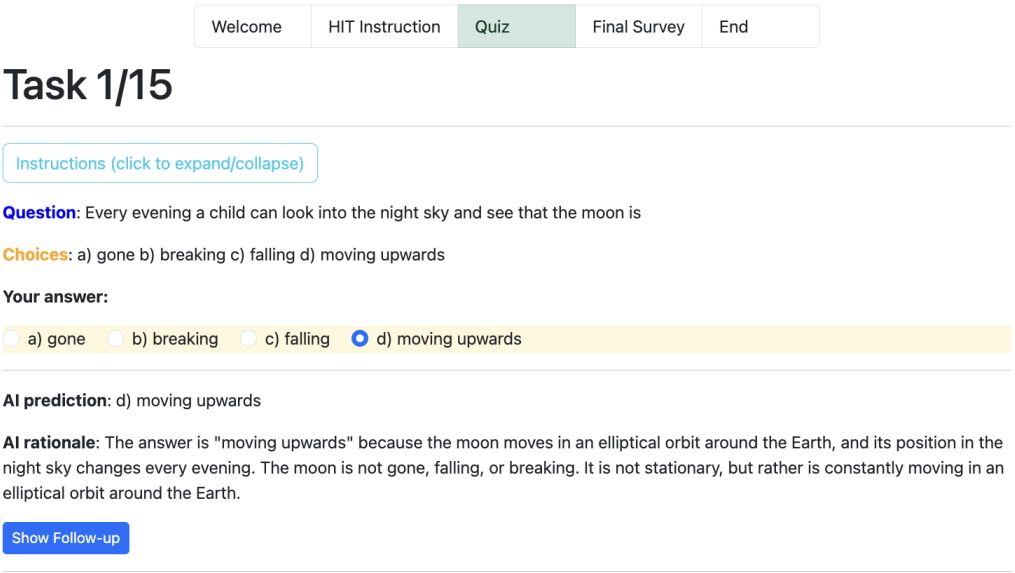


Figure 12: Once the participant selects an answer, they are shown the KIT model prediction and the LLM-generated rationale.

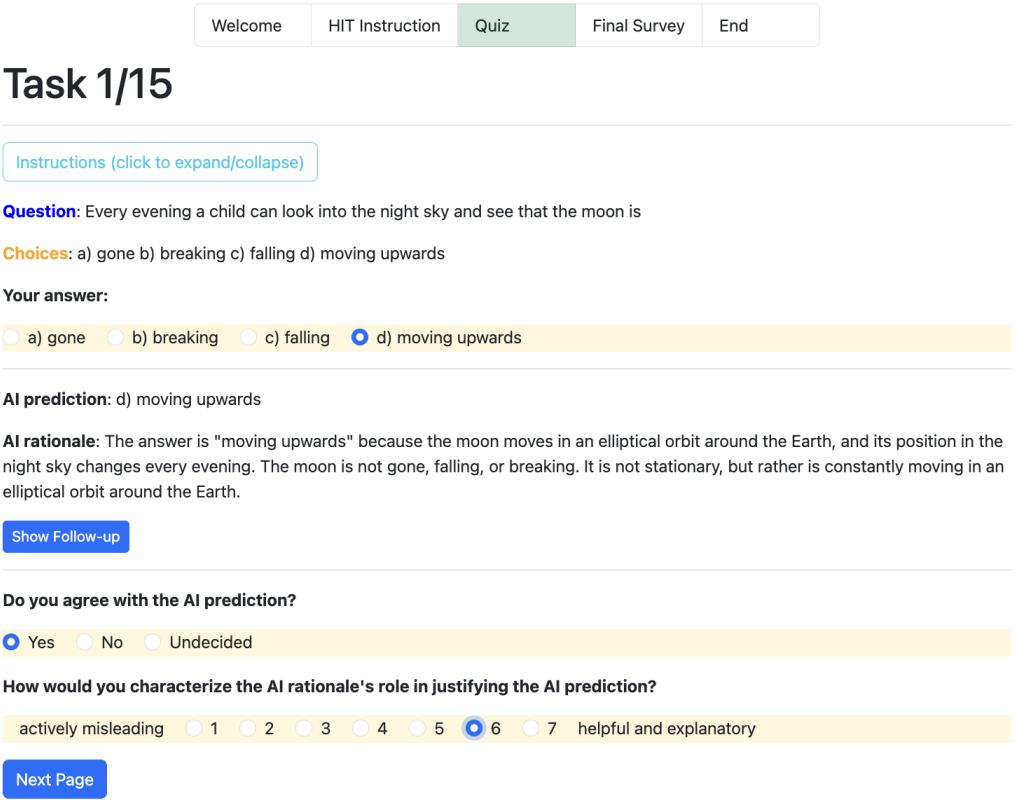


Figure 13: Collecting immediate participant feedback for a task.

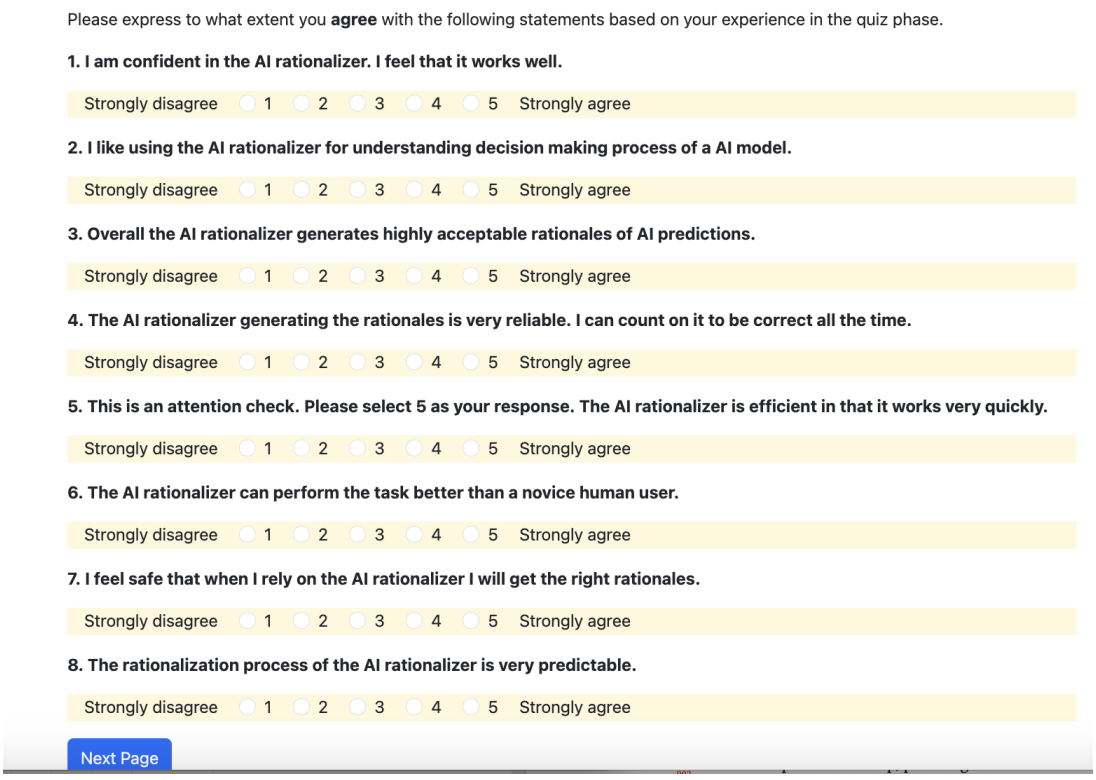


Figure 14: Trust scale-based survey of participant experience.