

# Digital Socrates: Evaluating LLMs through Explanation Critiques

Yuling Gu, Oyvind Tafjord, Peter Clark  
Allen Institute for AI, Seattle, WA  
{yulingg, oyvindt, peterc}@allenai.org

## Abstract

While LLMs can provide reasoned explanations along with their answers, the nature and quality of those explanations are still poorly understood. In response, our goal is to define a detailed way of characterizing the explanation capabilities of modern models and to create a nuanced, interpretable explanation evaluation tool that can generate such characterizations automatically, without relying on expensive API calls or human annotations. Our approach is to (a) define the new task of *explanation critiquing* - identifying and categorizing any main flaw in an explanation and providing suggestions to address the flaw, (b) create a sizeable, human-verified dataset for this task, and (c) train an open-source, automatic critique model (called Digital Socrates) using this data. Through quantitative and qualitative analysis, we demonstrate how Digital Socrates is useful for revealing insights about student models by examining their reasoning chains, and how it can provide high-quality, nuanced, automatic evaluation of those model explanations for the first time. Digital Socrates thus fills an important gap in evaluation tools for understanding and improving the explanation behavior of models.

## 1 Introduction

Large language models (LLMs) have demonstrated promising end-task performance on a range of tasks (OpenAI, 2023; Bubeck et al., 2023). These models, given their text-generation abilities, can also be prompted (Wei et al., 2022; Kojima et al., 2022) or trained (Rajani et al., 2019; Li et al., 2023; Tafjord and Clark, 2021; Gu et al., 2022) to externalize their reasoning<sup>1</sup> as a window into their reasoning capabilities. Despite promising end-task performance,

<sup>1</sup>A model-generated chain of reasoning can be a free-form explanation or a series of intermediate steps. We use this as a window into the model’s ability to create systematic arguments, but without making any claims about how LLMs reason internally (Shanahan, 2022; Andreas, 2022).

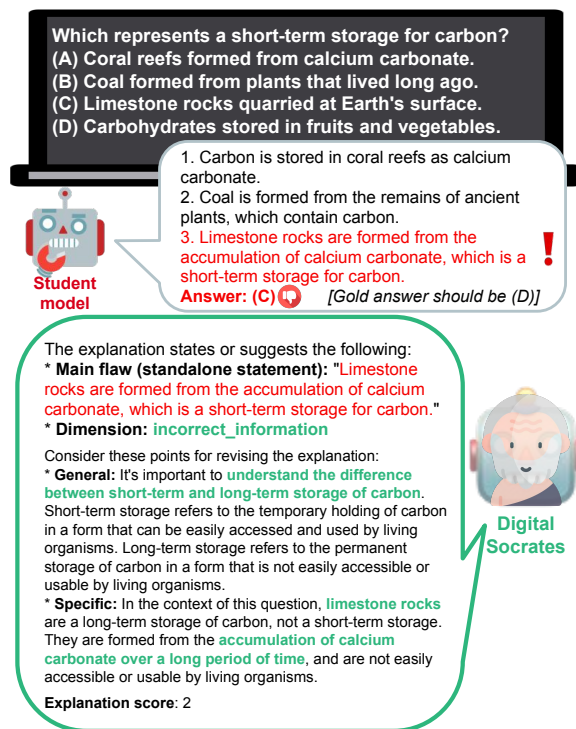


Figure 1: Given a multiple-choice question (together with the answer options and correct answer), as well as a model-generated reasoning chain and answer, our system Digital Socrates gives a critique of the model-generated explanation. In its critiques, Digital Socrates provides localized feedback on where and why reasoning chains are flawed (focusing on the main flaw, if any), accompanied by general and fine-grained suggestions to address the identified flaw, providing nuance and interpretability to the critiques.

examining LLMs’ reasoning chains reveals gaps in the correctness of their factual knowledge and the coherence of their reasoning (Radhakrishnan et al., 2023; Lanham et al., 2023; Ye and Durrett, 2022; Cohen et al., 2023). Such efforts delving deeper into the quality of model-generated intermediate reasoning chains enable us to advance our understanding of LLMs’ strengths and weaknesses in different tasks beyond measuring their performance on standard benchmarks.

In existing NLP works, however, judging the quality of LLMs’ intermediate reasoning chains is met with several challenges. Early practices for determining the quality of such intermediate generations include (1) reporting end-task accuracy and (2) sampling a subset to perform human annotations on e.g., Liu et al. (2022); Gu et al. (2022); Li et al. (2023). Using end-task performance as a proxy for the quality of intermediate reasoning can be problematic as LLMs’ final answers can be unfaithful to the generated intermediate reasoning (Radhakrishnan et al., 2023; Lanham et al., 2023), whereas relying on human annotations is labor-intensive and expensive.

To tackle such challenges, several automatic measurements have been proposed. One category focuses on evaluating model-generated text based on a given reference text  $y$  (Saunders et al., 2022; Kim et al., 2023; Xu et al., 2023). Another proposes numerical metrics like ROSCOE (Golovneva et al., 2023) and ReCEval scores (Prasad et al., 2023) as a summary of reasoning quality. Others build models that generate free-form critiques and refinement suggestions e.g., Wang et al. (2023). Our work builds upon these previous efforts, introducing a way of automatically evaluating reasoning chains that (1) focuses on the intrinsic quality of the reasoning chains, moving away from the reliance on comparing to any reference reasoning chain; (2) localizes where the reasoning went wrong and provides interpretable feedback on why that part of the reasoning chain should be revised; and (3) uses a semi-structured format useful for gaining both quantitative and qualitative insights about the reasoning chain quality.

To operationalize this, our approach and contributions are thus as follows:

- We define the task of *explanation critiquing*.
- We create DS Critique Bank, a sizeable, human-verified dataset for the task, both to train critique models and to compare against explanation capabilities of future models.
- We train and release a high-performing, open-source critique model, Digital Socrates, that does not rely on expensive API calls or human annotations.
- We demonstrate the usefulness of Digital Socrates critiques.

By providing high-quality, nuanced automatic evaluation of explanations, Digital Socrates fills an important gap in evaluation tools for the community. We make our dataset and model pub-

licly available at <https://allenai.org/data/digital-socrates>.

## 2 Related Work

**LLM Evaluation.** End-task performance on popular tasks is a standard way of evaluating models in NLP. For instance, Pythia (Biderman et al., 2023) and Llama-2 (Touvron et al., 2023) mostly focus on reporting accuracy on popular tasks like WSC (Levesque et al., 2011), ARC (Clark et al., 2018) or benchmarks like MMLU (Hendrycks et al., 2021), and use that to guide model development. We evaluate explanations as an additional dimension.

**Explanations and Rationales.** Free-form explanations and rationales have emerged as a promising approach to materialize models’ reasoning chains (Wei et al., 2022; Kojima et al., 2022; Schwartz et al., 2020; Gu et al., 2022; Chen et al., 2023; Wiegrefe et al., 2021). Despite LLMs’ impressive accuracy on a range of tasks (OpenAI, 2023; Brown et al., 2020), examining model-generated explanations surfaces issues such as the utility of such explanations to humans (Joshi et al., 2023), the correctness of models’ factual knowledge, and the coherence of their reasoning (Radhakrishnan et al., 2023; Lanham et al., 2023; Ye and Durrett, 2022; Cohen et al., 2023). We propose the *explanation critiquing* task as a systematic way to examine models’ explanation quality.

**Model-generated Feedback/Critique.** Prior works like Qi et al. (2023) and Madaan et al. (2023) focused on utilizing model-generated feedback to improve task performance. In terms of using automatic feedback for analyzing models, ROSCOE (Golovneva et al., 2023) provides numeric scores covering semantic alignment, semantic similarity, logical influence, and language coherence; ReCEval scores (Prasad et al., 2023) focus on the correctness and informativeness of reasoning chains. Continuing this line of work in a more general and flexible manner, our *explanation critiquing* task involves nuanced and interpretable critiques beyond numeric scores. Different from works like Wang et al. (2023) which suggest refinements in free-form text, our critique design is informed by principles of Socratic questioning (Brunschwig et al., 2003; Paul and Elder, 2006). Further, extending efforts on critiquing model generations utilizing reference text by Saunders et al. (2022); Kim et al. (2023); Xu et al. (2023), we move away from any reliance on reference texts.

### 3 Explanation Critiquing: Task Design

We formalize the *explanation critiquing* task in this section. Then in the rest of the paper, we describe how we collect a dataset for this task, build a small and high-performing model for this task, and finally apply the model to reveal interesting insights about QA models in an automatic manner.

Given a question  $Q$ , along with a model-generated explanation  $E_{pred}$  and answer  $A_{pred}$ , the task involves giving a **critique of the model-generated explanation**, expressed as a 5-tuple:

$$critique(Q, E_{pred}, A_{pred}) = \langle f_{loc}, f_{dim}, s_{gen}, s_{spec}, E_{SC} \rangle$$

where  $f_{loc}$  is a natural language (NL) statement that localizes and describes the most significant flaw (if any) in the explanation,  $f_{dim}$  categorizes  $f_{loc}$  into one of 8 categories of flaws described in Section 3.1 (or “None”),  $s_{gen}$  and  $s_{spec}$  are general and specific NL suggestions for correcting the flaw, and  $E_{SC}$  is a numeric rating (from 0 to 5) indicating the explanation quality. See example critique in Figure 1.

#### 3.1 Revealing Problems in the Explanation

We design the first component,  $f_{loc}$ , of the critique to identify the most significant flaw (if any) in the explanation, pinpointing explicitly which part of the explanation is problematic. The flaw should be formulated as a standalone erroneous statement, that can be understood without reference to the full question or explanation.<sup>2</sup>

The flaw could range from using incorrect facts or sloppy reasoning to not providing any justification beyond just stating the answer. Informed by the systematic and disciplined method of Socratic questioning (Brunschwig et al., 2003), we target the critique to serve as a way of questioning the given explanation “to get to the truth of things, to open up issues and problems, to uncover assumptions, to analyze concepts ... and to follow out logical implications of thought.” (Paul and Elder, 2006) Specifically, our critique focuses on flaws along the 8 dimensions ( $f_{dim}$ ) detailed in Table 1. These dimensions have been chosen to cover the different types of Socratic questions enumerated in Paul and Elder (2006) to rigorously examine if the explanation demonstrates an understanding of the question, uses valid reasoning and evidence,

considers appropriate viewpoints and perspective, and derives logical implications and consequences.

#### 3.2 Suggestions toward Truth and Accuracy

The critique also contains general and specific suggestions to ensure that each flaw identified is justified with a direction for improvement in mind, rather than just being overly critical. The general suggestion,  $s_{gen}$ , is a statement that addresses a likely misconception or confusion underlying the flaw. This should not directly give out a particular answer to the question. Such information could then be used in future efforts to improve broad failure types in the explanation model. The specific suggestion,  $s_{spec}$ , is a more targeted statement to consider when revising the explanation, as a guide towards the right reasoning chain (to the right answer) for this particular question.

#### 3.3 Explanation Score

The *explanation critiquing* task also involves providing a quantitative metric on the explanation quality. The explanation score  $E_{SC}$  is on a scale from 0 to 5 as summarized in Table 2.

### 4 Dataset of Explanation Critiques

We introduce DS Critique Bank,<sup>3</sup> a critique dataset where each instance comprises a multiple-choice question (together with answer options and correct answer), a model-generated explanation and answer (using a variety of different models), a critique of the model-generated explanation, as well as (any) human annotations collected on that instance. DS Critique Bank focuses on questions requiring reasoning, in particular science and commonsense reasoning (Section 4.1). The explanations are from different models in popular explanation styles (Section 4.2). The generation of critiques as well as the process for crowdsourced and expert annotations are described in Section 4.3.

#### 4.1 Tasks

To obtain questions that require reasoning about basic science concepts, we sample from the AI2 Reasoning Challenge (ARC) dataset (Clark et al., 2018), which contains natural, grade-school level, multiple-choice science questions. For commonsense reasoning questions, we sample from the RAINBOW (Lourie et al., 2021) suite of tasks comprising 6 multiple-choice question-answering

<sup>2</sup>More details on format in Appendix C.1.

<sup>3</sup>Digital Socrates’ Critique Bank.

Critique dimension	Description of flaw in the explanation
misunderstanding	apparent misunderstanding of the original question or answer choices
lack justification	just stating the final answer without any proper or informative justification
incorrect information	stating incorrect fact(s) or knowledge
missing information	missing a crucial fact, knowledge, or perspective that should be considered
incorrect reasoning	an incorrect leap in the reasoning
incomplete reasoning	the reasoning doesn't lead all the way to the answer
inconsistent answer	the answer doesn't match the conclusion of the explanation
irrelevant	using irrelevant or redundant fact(s), knowledge, or reasoning

Table 1: Critique dimensions in our *explanation critiquing* task. These dimensions cover the different types of Socratic questions (Paul and Elder, 2006) to guide rigorous examination of explanation quality.

$E_{SC}$	Description of explanation quality
0	Completely wrong
1	Has something useful relevant to the question
2	Makes valid points, but mostly incorrect
3	Many correct elements, but with significant flaws
4	Mostly correct, with minor flaw
5	Completely correct

Table 2: Descriptions for different explanation scores  $E_{SC}$  in the *explanation critiquing* task, used to rate explanation quality on a scale from 0 to 5.

Partition	#Qs	#Critiques	#Annos
DSCB-train-silver	3240	3240	0
DSCB-train-crowd-anno	270	3240	4040
DSCB-train-expert	31	198	198
DSCB-dev-crowd-anno	550	19800	2430

Table 3: DS Critique Bank partitions provide breadth across different questions (DSCB-train-silver) as well as depth in comprehensively covering different student models and explanation styles (DSCB-train-crowd-anno). It contains both crowdsourced (DSCB-train-crowd-anno) and expert (DSCB-train-expert) annotations. The dev set is also released (DSCB-dev-crowd-anno). More details are in Appendix D.

datasets –  $\alpha$ NLI (Bhagavatula et al., 2020), CosmosQA (Huang et al., 2019), HellaSwag (Zellers et al., 2019), Physical IQa (Bisk et al., 2020), Social IQa (Sap et al., 2019) and WinoGrande (Sakaguchi et al., 2020). For evaluation purposes, we create a separate “dev” (evaluation) partition containing additional samples from ARC and RAINBOW, plus samples from two additional datasets (i.e., OOD wrt. training), namely OpenBookQA (OBQA) (Mihaylov et al., 2018) and CommonsenseQA (CSQA) (Talmor et al., 2019). Table 3 summarizes the main statistics of DS Critique Bank, with more details on sample sizes and mixes in Appendix D.

## 4.2 Explanations from Models

### 4.2.1 Explanation Styles

We prompt models to generate an explanation along with their answers when performing QA tasks.

We use three different prompts (see Appendix G) which cover explanation styles in the spirit of Zero-shot Chain-of-Thought reasoning (CoT) (Kojima et al., 2022), Few-shot CoT (Wei et al., 2022), and reasoning steps (Dalvi et al., 2021).

### 4.2.2 Explanation (“Student”) Models

We use the following models to generate such explanations: GPT-4-0613 (OpenAI, 2023), GPT-3.5-turbo-0613 (Brown et al., 2020), Llama2-7B-chat, and Llama2-70B-chat (Touvron et al., 2023). This covers strong proprietary models and open models of different strengths. For the rest of the paper, we will refer to these models whose outputs are to be critiqued as **student models**.

## 4.3 Explanation Critiques

We elicit seed explanation critique data from GPT-4 (OpenAI, 2023)<sup>4</sup> and perform expert<sup>5</sup> and crowdsourced annotations on our training data (see Table 3). The expert annotation involves (i) rating the GPT-4 generated critique (according to scales in Table 4) and (ii) revising bad critiques to make them good ones. The crowdsourced annotations involve (i) following the guidelines for *explanation critiquing* to identify applicable critique dimensions of any significant flaws in the explanation, as well as rating the explanation quality; (ii) rating the GPT-4 generated critique.<sup>6</sup> We release expert-revised critiques, crowdsourced explanation scores and flaw dimensions, as well as all critique rating scores as part of DS Critique Bank.

As part of our evaluation, we also crowdsourced annotations of dev set critiques generated by our trained critique models (described below in Section 5). These annotations are also included in the dataset release, and described in Section 6.

<sup>4</sup>Latest model at experimentation, GPT-4-0613.

<sup>5</sup>Annotated by two of the authors.

<sup>6</sup>See Appendix B for more details.



Critique score	Description of critique quality
0	Bad quality critique
1	Mostly bad quality critique, but makes some useful point(s)
2	Mostly good quality, helpful critique, but can be improved
3	Very good quality critique

Table 4: Descriptions for different critique quality scores. Generated critiques are scored on a scale of 0 to 3, where 0 indicates the critique is very bad and 3 means that the critique is of very good quality. To encourage the annotators to make an overall decision about acceptability (2+) of the critique, we avoided including an “in-between” category (e.g., 1.5).

#### 4.4 Statistics

The final dataset contains 26,478 entries of nuanced explanation critiques for 4,091 different questions in the science and commonsense domains. Table 3 summarizes the main statistics of our DS Critique Bank, with more details in Table 7 in Appendix D. To the best of our knowledge, this is the first dataset of its kind on *explanation critiquing*, covering nuanced and interpretable (user comprehensible) critiques on different models’ explanations and in different explanation styles. Its roles are to both train automated critique models, and to compare new models’ explanation capabilities against.

### 5 Our Model: Digital Socrates

In addition to probing GPT-4 (not fine-tuned) as a critique model, we also train two smaller models, DS-7B and DS-13B,<sup>7</sup> to explore whether *smaller* language models can be fine-tuned to perform the task of *explanation critiquing* well.

#### 5.1 Training Data

We use the train partitions of DS Critique Bank where we apply two filters: (i) down-sample critiques finding no flaws so they make up at most 50%; (ii) only keep critiques that are judged as good quality by annotators (with a score of 2 or 3).

We order the training set by degree of annotation, arriving at the following curriculum of instances:

- 2032 from DSCB-train-silver (1 epoch)
- 1640 from DSCB-train-crowd-anno (1 epoch)
- 198 from DSCB-train-expert (2 epochs)

<sup>7</sup>We use DS as the short form for Digital Socrates.

#### 5.2 Training

We fine-tune two critique models DS-7B and DS-13B starting from Llama2-7B-chat and Llama2-13B-chat (Touvron et al., 2023) respectively. First, we pre-fine-tune on a set of about 50k training questions from ARC and RAINBOW, doing a simple zero-shot question-answering task. Then we further fine-tune on the curriculum of increasing critique quality as indicated above (Section 5.1). See Appendix I for more details.

### 6 Experiments

We performed experiments and analyses to answer three questions: (1) What types of insights can the *explanation critiquing* task reveal about student models? (2) How good is GPT-4 at *automatic critiquing*? (3) How well can a *smaller, trained model* (thus avoiding expensive API calls) perform automated critiquing?

To answer the first question, we analyze the human-annotated training data to identify what these “ideal” critiques can tell us about the student models (Section 7.1). To answer the second and third questions (i.e., can we generate such critiques automatically?), we first ran our three critique models (GPT-4, and the trained models DS-7B, DS-13B) to generate critiques of explanations in the dev partition of our dataset (DSCB-dev-crowd-anno). We then obtained human annotations on a subset of these, namely the three critiques on student model explanations for 270 dev questions (drawn from all 10 datasets in dev), i.e., 810 annotated critiques (each done by 3 annotators).<sup>8</sup> Finally we use this data to evaluate each critique model (Sections 7.2 and 7.3).

### 7 Results and Analysis

#### 7.1 What can *explanation critiquing* inform us about student models?

In this section, we discuss the types of insights we can get from the *explanation critiquing* task, based on human-annotated critique data from the DSCB-train-crowd-anno partition of DS Critique Bank. In subsequent sections, we show that this carries over to model-generated critiques as well.

We analyze the distribution of explanation scores  $E_{SC}$  separately for questions where the student model answers correctly vs incorrectly (Figure 2).

<sup>8</sup>A total of 32 annotators contributed to the whole annotation effort. Further details are in Appendix D.

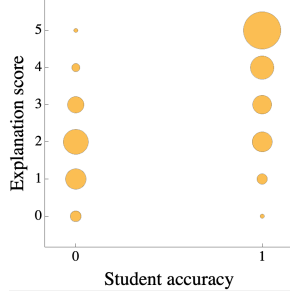


Figure 2: In student models, (human-annotated) explanation scores  $E_{SC}$  vary greatly within cases where models get the answer right (accuracy = 1) or wrong (accuracy = 0). Even when a model gets the answer correct, its reasoning chain can contain varying degrees of flaws. On the other hand, when a model is incorrect in its answer, it could still make some valid points.

Even when a model gets the answer correct, it can give a flawed reasoning chain, ranging from minor flaws ( $E_{SC} = 4$ ) to something completely wrong ( $E_{SC} = 0$ ). Conversely, when a model is incorrect in its answer, it could still make some valid points ( $E_{SC} = 2$ ), thus not all incorrect final answers should be viewed the same as a complete lack of understanding required for the question. Such trends are **seen across models**, regardless of model size and type (see Figure 8 in Appendix E.1).

The *explanation critiquing* task also provides the dimension of the most significant flaw, giving much explainability and interpretability. Using GPT-4 generated critiques rated as good quality by crowdworkers (critique score 2 or 3), Figure 3 summarizes the overall fractions of different significant flaws in models’ reasoning chains. When models answer questions correctly, there can still be cases where the reasoning chain has flaws (13% across all student models) such as incorrect information. When the model is incorrect, there is a broad spectrum of flaws, such as misunderstanding the question or answer options, incorrect information or reasoning, and reasoning inconsistent with the given answer.

As a case study, we compare two student models, GPT-3.5 and Llama2-70B, on the Science domain (Figure 4). The models have comparable accuracy, with a slightly lower explanation score  $E_{sc}$  for the latter, hinting at shallower understanding. The pie charts showing the flaw dimensions provide an efficient snapshot of the models’ weaknesses. For incorrect answers, we see that Llama2-70B has a higher rate of incorrect information, which could potentially be mitigated by information retrieval or otherwise improving factuality. GPT-3.5 instead

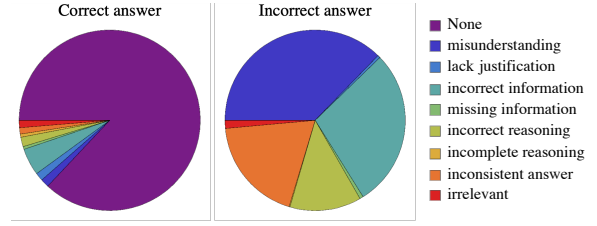


Figure 3: The pie charts show distributions of explanation flaws across all student models. Even when models get the answer correct, they may still make errors in their reasoning chain (left). When models answer incorrectly, *explanation critiquing* helps in categorizing and diagnosing errors in the reasoning chain (right).

has a notably large fraction of inconsistent answer, where the answer does not match the reasoning in the explanation. Future work improving GPT-3.5 could potentially experiment with self-reflection, asking if the answer follows from the explanation. We extend this picture across all student models and datasets, discussing further model and domain-specific insights in Appendix E.2

Student model	Acc	$E_{SC}$	Dimensions acc=1	Dimensions acc=0	$E_{SC}$ acc=1	$E_{SC}$ acc=0
GPT-3.5	0.84	4.43			4.87	2.15
Llama2-70B	0.81	4.18			4.72	1.9

Figure 4: GPT-3.5 and Llama-2-70B student models achieve comparable *Acc* on Science datasets, with the latter having slightly lower  $E_{SC}$ . They also show different patterns in their explanations flaws, e.g., in the amount of incorrect information vs inconsistent answer.

Our *explanation critiquing* task formulation delivers even more nuance and interpretability by localizing the flaws.  $f_{loc}$  pinpoints flawed parts of model-generated explanations, like “Ferromagnetic (C) materials are not typically used for paper clips.”<sup>9</sup> Beyond looking at the dimensions of explanation flaws to note that incorrect information is frequent in Llama2-70B model’s explanations compared to GPT-3.5, localizing the flaws further informs us about topics in which the model has incorrect information. Having a **list of such flaws** for each model could better **facilitate finding patterns in models’ gaps in knowledge** and reasoning.

The *explanation critiquing* task also provides

<sup>9</sup>Example of incorrect information in Llama2-70B. More examples of identified flaws are in Table 9 in Appendix E.3.

Student model	Dataset	General feedback examples (acc=0)
GPT-4	Science	Remember that the state of matter is indeed a physical property that can distinguish between different types of elements. Metals are generally solid at room temperature, while non-metals can be gases, liquids, or solids.
GPT-3.5	Science	Remember that metals are excellent conductors of electricity due to their free electrons.
Llama2-70B	Science	Always remember that in multicellular organisms, the hierarchy of complexity begins with the cell, then progresses to tissues, organs, and finally organ systems.
Llama2-7B	Science	When conducting an experiment, it's crucial to control for as many variables as possible to ensure that any observed differences are due to the factor you're testing (in this case, the type of plant food), not other factors.
GPT-4	Commonsense	Always ensure you understand the question and the answer choices. The question is not asking what Skylar will definitely do next, but what someone might want Skylar to do next.
GPT-3.5	Commonsense	Ensure that your explanation aligns with the answer you have chosen. It's important to be consistent in your reasoning and conclusion.
Llama2-70B	Commonsense	Be careful with interpreting the language used in the question and answer choices. Sometimes, phrases may not be literal but idiomatic or figurative.
Llama2-7B	Commonsense	Sensitivity to spice typically means a person can't handle much spice, not that they add more spice to their food.

Table 5: Examples of general suggestions sampled from human-verified critiques in DSCB-train-crowd-anno. The *explanation critiquing* task includes providing a suggestion that addresses general misconceptions or confusion underlying mistakes that models make under each dimension.

general and specific suggestions to correct the flaw. For instance, providing correct information like “Ferromagnetic materials, such as iron, nickel, and cobalt, ... in a variety of applications, including holding sheets of paper together.”<sup>10</sup> The **general feedback** (Table 5) could then offer **directions toward streamlining model improvement** or serve as **useful retrieval corpus**, while the specific feedback helps to correct reasoning for each instance.

With *explanation critiquing*, as informed by the principles of Socratic questioning, the nuanced and interpretable critiques offer various components useful as a systematic, rigorous approach for evaluating model-generated explanations. Such fine-grained information allows for effective uncovering of specific errors associated with models, which can also be used to improve models in the future.

## 7.2 To what extent can GPT-4 perform the task of *explanation critiquing*?

While *explanation critiquing* allows for systematic, insightful analysis of models, obtaining human-corrected or even human-verified explanation critiques can be expensive and time-consuming. We analyze if raw critiques generated by GPT-4 can be used to reveal the same trends,<sup>11</sup> using the DS Critique Bank dev partition (DSCB-dev-crowd-anno).

**Human judgments of critique quality** (Table 6) show that 92% of critiques generated by GPT-4 are

<sup>10</sup>Example of a general suggestion to Llama2-70B’s explanation.

<sup>11</sup>Compared to obtaining human annotations, API calls are still relatively cheaper in time and monetary cost.

Critique model	#Qs	Rated good	Rated good*	Dimension overlap	$E_{SC}$ match
GPT-4	270	0.92	0.96	0.81	0.88
DS-13B	270	0.82	0.92	0.84	0.87
DS-7B	270	0.75	0.89	0.84	0.88

Table 6: Comparison of critique quality from GPT-4, DS-7B, and DS-13B, as rated by crowdworkers. We compare critiques generated by each critique model in terms of the percentage that is rated good (critique score 2 or 3) by workers, how often the flaw dimension identified overlaps with human annotations, and explanation score ( $E_{SC}$ , cases where model-generated and human-rated explanation score is within 1). The “Rated good\*” column extrapolates to the actual distribution of the full dataset by accounting for the large fraction (57%) of examples with None critiques from all 3 models.

rated mostly good or very good quality. This number extrapolates to 96% when we account for the fact that a large portion of the full dev set consists of explanations that received a “None” critique (such instances all rated correct in crowdsourcing) from all 3 critique models. GPT-4 critiques also **agree well with crowdsourced *explanation critiquing***. In 81% of the cases, the GPT-4 critique is along a dimension that overlaps with annotators’ identified flaw dimensions, and 88% of GPT-4 critiques give an explanation score within 1 of that given by annotators on the 6-point scale (Table 2), thus providing similar signals in analysis. We also observe very high correlation between annotators and GPT-4 for explanations scores, with Pearson’s correlation coefficient  $r = 0.73$ . These metrics suggest that

analysis done directly using the GPT-4 generated critiques strongly agrees with human-verified data.

### 7.3 To what extent can our smaller, trained models (Digital Socrates 7B, 13B) perform the task of *explanation critiquing*?

We evaluate the critique quality of DS-7B and DS-13B, both in terms of overall critique quality, as well as flaw dimension overlap and explanation score match compared to human annotations.

Based on **human judgments of critique quality** (Table 6), the vast majority of critiques generated by our DS models are rated good or very good, at 82% for DS-13B and 75% for DS-7B. These numbers increase to 92% and 89% when extrapolating to the full distribution.<sup>12</sup> Our DS models' critiques **highly agree with human explanation critiquing**. For both DS-7B and DS-13B, 84% of the time, the generated critique is along a dimension that overlaps with annotators' identified flaws, comparing favorably to the 81% dimension overlap for GPT-4 critiques. This suggests our DS models are even more closely aligned to human judgments in terms of identifying the flaw dimension in explanations, potentially due to the effect of human-annotated data in their training. The percentage of critiques that give an explanation score within 1 of that given by annotators, is 88% for DS-7B ( $r = 0.76$ ) and 87% for DS-13B ( $r = 0.73$ ), which is comparable to that for GPT-4 (88%).

## 8 Further Analysis on Digital Socrates

Having established the validity of the automatic critique models, especially GPT-4 and DS-13B, we use them to analyze all instances in our dev set.

**DS-13B critiques reveal similar trends on student models compared to GPT-4.** Figure 6 presents a representative example comparing the flaw dimension pie charts across critique models. DS-13B and GPT-4 models reveal very similar trends in student models' flaw dimensions, while the DS-7B is slightly less diverse in its critiques. Appendix E.5 provides the full comparison.

**Further, DS-13B critiques reveal similar trends on student models compared to human-verified critiques.** Compared to trends identified in Table 8 using human-verified data, DS-13B identifies similar characterizations of models (Figure 9). For instance, comparing when models answer the

<sup>12</sup>Despite these models being orders of magnitude smaller than GPT-4.

**Question:** When Dennis operates his lawnmower, he notices the engine makes a lot of noise. He also notices that the engine gets very hot. Which best describes the heat and noise generated from the lawnmower? (A) a change in phase (B) thermal expansion (C) an increase in entropy (D) mechanical advantage

**Explanation (GPT-3.5):** 1) The question states that the lawnmower engine makes a lot of noise. 2) The question states that the lawnmower engine gets very hot. 3) Noise and heat are both forms of energy. 4) The noise and heat generated from the lawnmower are a result of the conversion of energy from the fuel to mechanical energy.

**Answer (GPT-3.5):** (D) [incorrect, should be (C)]

#### DS-13B Critique:

The explanation states or suggests the following:

\* **Main flaw** (standalone statement): "The noise and heat generated from the lawnmower are a result of the conversion of energy from the fuel to mechanical energy."

\* **Dimension:** incorrect\_information

Consider these points for revising the explanation:

\* **General:** It's important to understand the difference between the different types of energy. Mechanical energy is the energy of motion, while thermal energy is the energy of heat.

\* **Specific:** In the case of the lawnmower, the noise and heat are not a result of the conversion of energy from the fuel to mechanical energy. The noise is a result of the vibration of the engine, while the heat is a result of the friction and combustion of the fuel.

**Explanation score:** 2

Figure 5: Sample explanation critique from DS-13B.

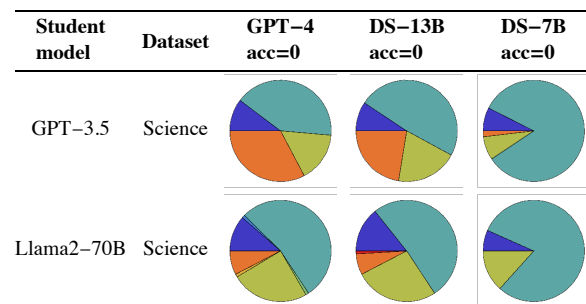


Figure 6: Sample comparison of insights on student models based on critiques from GPT-4, DS-7B, and DS-13B using the full dev partition (DSCB-dev-crowd-anno). The distribution of explanation flaws as identified by DS-13B is particularly close to that by GPT-4.

questions incorrectly across dataset types, in the case of reasoning commonsense reasoning, student models tend to show a misunderstanding of the question or answer choices, whereas incorrect information is more prevalent as a flaw for science datasets. Model-specific insights are also similar. E.g., Llama2-70B's most common flaw when performing scientific reasoning concerns incorrect information. Whereas in GPT-3.5, that flaw is somewhat less frequent but having a reasoning chain inconsistent with the answer is also of concern.

### DS-13B reveals insightful trends on student



**models.** We qualitatively examine flaws localized and suggestions offered by DS-13B. Similar to GPT-4, DS-13B provides high-quality critiques that accurately localize main flaws in models’ explanations. Figure 5 shows an example of how DS-13B accurately identifies the somewhat subtle error “The noise and heat generated from the lawnmower are a result of the conversion of energy from the fuel to mechanical energy.” as incorrect information, providing the accurate feedback of “... Mechanical energy is the energy of motion, while thermal energy is the energy of heat.”

In Appendix F, we cover other top common flaw dimension examples from DS-13B. We further demonstrate, in Appendix E.4, how comprehensive results for applying DS-13B on all 10 datasets of the DS Critique Bank dev set, across all 4 student models, reveal a rich diversity of behavior.

## 9 Conclusion

Despite advances in LLMs, the nature and quality of their explanations are still poorly understood. To address this, we introduced the task of *explanation critiquing* which provides quantitative and qualitative insights into LLM performance, beyond answer accuracy. Our analysis highlights the value of closely examining model-generated explanations for a deeper understanding of models’ capabilities. We create DS Critique Bank, a dataset of 26,478 nuanced and interpretable critiques under the *explanation critiquing* task formulation, along with crowdsourced and expert annotations. Our Digital Socrates model (DS-13B), is a capable open critique model comparable in performance to GPT-4 and closely aligned with human annotators in the *explanation critiquing* task. We also showed how Digital Socrates can, for the first time, provide rich analyses and insights across a range of student models and datasets, without relying on expensive API calls or human annotation. This work fills an important gap in evaluation tools for understanding and improving the explanation behavior of models.

## Acknowledgements

We thank the anonymous reviewers, as well as members of the Aristo team at AI2 for their insightful discussions and valuable feedback on an earlier draft.

## Limitations

We have demonstrated the effectiveness of Digital Socrates to provide evaluation and insights into a model’s capabilities going beyond model accuracy. While exciting, there are limitations in the current work and opportunities for the future.

Our approach is designed for examining model-generated reasoning chains involving natural language explanations or reasoning steps, focusing on focus on linguistic reasoning (in broad domains of commonsense and science). Building on a better understanding of such reasoning chains, we hope our work can inspire future efforts towards extending the approach in various directions, for instance, for application to longer text e.g., model-generated summaries or stories, as well as symbol manipulation reasoning such as arithmetic, and formal logic. We encourage other researchers to build upon our work in other domains, and even evaluate our existing Digital Socrates models on other reasoning types since the prompt to the models and inherent Socratic questioning principles are domain agnostic. It is thus possible the DS models would apply to such reasoning styles also, and would be an interesting avenue for future work.

Another interesting direction for future work is investigating how well different student models can respond to the feedback, and to what extent feedback from  $s_{gen}$  can generalize to help student models with related questions.

The current Digital Socrates models have to rely on internal knowledge, so like any LLM can be subject to hallucination of factual knowledge. We encourage the use of Digital Socrates models for research purposes only and for researchers to build on them responsibly.

## Ethics Statement

All annotators that participated in the data collection process have been anonymized. The only personal information we collect is the worker IDs from Amazon Mechanical Turk, which we will not release. No personally identifiable information is contained in our dataset or otherwise released. We took great care to pay fair wages, and were responsive to feedback and questions throughout the data collection process. This study involves the use of large-scale language models. Like any other large-scale language model, despite the best intentions, there is a risk of our models producing biased or offensive statements as part of the critique text. We

release our models for research purposes only.

## References

- Jacob Andreas. 2022. [Language models as agent models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. [Abductive commonsense reasoning](#). In *ICLR*.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#). *CoRR*, arXiv:2304.01373.
- Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. [PIQA: Reasoning about physical commonsense in natural language](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Jacques Brunschwig, Geoffrey Ernest, Richard Lloyd, and Pierre Pellegrin. 2003. *A Guide to Greek Thought: Major Figures and Trends*. Harvard University Press.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with GPT-4](#). *CoRR*, arXiv:2303.12712.
- Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. [REV: Information-theoretic evaluation of free-text rationales](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030, Toronto, Canada. Association for Computational Linguistics.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? Try ARC, the AI2 reasoning challenge](#). *CoRR*, arXiv:1803.05457.
- Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. [LM vs LM: Detecting factual errors via cross examination](#). In *EMNLP*.
- Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. [Explaining answers with entailment trees](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *NeurIPS*.
- Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. [ROSCOE: A suite of metrics for scoring step-by-step reasoning](#). In *The Eleventh International Conference on Learning Representations*.
- Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022. [DREAM: Improving situational QA by first elaborating the situation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *ICLR*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-rank adaptation of large language models](#). In *ICLR*.
- Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Cosmos QA: Machine reading comprehension with contextual commonsense 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 2391–2401, Hong Kong, China. Association for Computational Linguistics.
- Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. [Are machine rationales \(not\) useful to humans? measuring and improving human utility of free-text rationales](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

- pages 7103–7128, Toronto, Canada. Association for Computational Linguistics.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. [Prometheus: Inducing fine-grained evaluation capability in language models](#). *CoRR*, arXiv:2310.08491.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiušė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. [Measuring faithfulness in chain-of-thought reasoning](#). *CoRR*, arXiv:2307.13702.
- Hector J. Levesque, Ernest Davis, and L. Morgenstern. 2011. [The Winograd schema challenge](#). In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.
- Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. [Symbolic chain-of-thought distillation: Small models can also “think” step-by-step](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2665–2679, Toronto, Canada. Association for Computational Linguistics.
- Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. 2022. [Rainier: Reinforced knowledge introspector for commonsense question answering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8938–8958, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nicholas Lourie, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2021. [UNICORN on RAINBOW: A universal commonsense reasoning model on a new multitask benchmark](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *NeurIPS*.
- 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](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- OpenAI. 2023. [GPT-4 technical report](#).
- Richard Paul and Linda Elder. 2006. *The Thinker’s Guide to Socratic Questioning: Based on Critical Thinking Concepts & Tools*. The Foundation for Critical Thinking, Dillon Beach, CA, USA.
- Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. [ReCEval: Evaluating reasoning chains via correctness and informativeness](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10066–10086, Singapore. Association for Computational Linguistics.
- Jingyuan Qi, Zhiyang Xu, Ying Shen, Minqian Liu, Di Jin, Qifan Wang, and Lifu Huang. 2023. [The art of SOCRATIC QUESTIONING: Recursive thinking with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4177–4199, Singapore. Association for Computational Linguistics.
- Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiušė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. [Question decomposition improves the faithfulness of model-generated reasoning](#). *CoRR*, arXiv:2307.11768.
- 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.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2020. [WinoGrande: An adversarial winograd schema challenge at scale](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8732–8740.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. [Social IQa: Commonsense reasoning about social interactions](#). 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.



- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. [Self-critiquing models for assisting human evaluators](#). *CoRR*, arXiv:2206.05802.
- Murray Shanahan. 2022. [Talking about large language models](#). *arXiv preprint*, arXiv:2212.03551.
- Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. [Unsupervised commonsense question answering with self-talk](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.
- Oyvind Tafjord and Peter Clark. 2021. [General-purpose question-answering with Macaw](#). *CoRR*, arXiv:2109.02593.
- 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.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *CoRR*, arXiv:2307.09288.
- Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O’Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. [Shepherd: A critic for language model generation](#). *CoRR*, arXiv:2308.04592.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Sarah Wiegrefe, Ana Marasović, and Noah A. Smith. 2021. [Measuring association between labels and free-text rationales](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. [InstructScore: Explainable text generation evaluation with finegrained feedback](#). *CoRR*, arXiv:2305.14282.
- Xi Ye and Greg Durrett. 2022. [The unreliability of explanations in few-shot prompting for textual reasoning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392. Curran Associates, Inc.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.



## A FAQs

### Q: What is the role of DS Critique Bank in future work?

There are three key roles:

- (a) Comparative analysis of different models' explanation capabilities: Given a user has a new model to test, they would run our dataset's test questions on their model, then use one of our Digital Socrates to critique the model's answers. It is important to run the same test questions in such tests so that the results can be compared with those generated by other models, if some kind of comparative analysis is needed (e.g., is  $GPT_n$  generating better explanations than  $GPT_{n-1}$ ?)
- (b) Creation of new Digital Socrates: As new models become available (e.g., new small-but-powerful models on the horizon), the dataset can be used to create new (and hopefully even stronger) Digital Socrates critiques models. The explanations in the dataset were generated by different models including strong proprietary models and open models of different strengths, and thus cover a broad variety of styles/techniques. By training Digital Socrates using this data, our model is designed to be able to robustly evaluate multiple styles of explanation.
- (c) General understanding of explanations: This is the first dataset containing detailed critiques of explanations, and as such could be useful to future researchers wanting to develop better explanation generation methods.

### Q: Is the dataset creation replicable in the future?

Yes! It's important to note that the dataset contains human-vetted GPT-4 critiques (of explanations from 4 different student models), not simply GPT-4 output. Thus the dataset is not a naive reflection of GPT-4, but of human judgments about what critiques are good and which are not - this is critically important. It would be easy to repeat this process using other models, along with vetting to expand the dataset (GPT-4 gave us the highest "hit rate" of good examples). So the process is easily repeatable, just like other AI datasets.

### Q: Why are the critiques interpretable?

The intended contrast here is between prior methods offering just a single score ("uninterpretable", in that the reason for the score is opaque), vs. the Digital Socrates, where the critiques tell you the category of the main error (along 8 dimensions guided by Socratic questioning principles), as well as actionable hints to improve the explanation, i.e., user comprehensible ("interpretable") reasons why the explanation is good or bad.

### Q: Why is the approach rigorous?

Digital Socrates critiques methodically step through several aspects of a critique (main flaw, dimension, correction, etc.) rather than just giving vague feedback like "The explanation is not detailed enough". Section 3 introduces the approach and details why it is systematic and methodical.

### Q: For rating critique quality, the gap between scores 1 ("mostly bad") and 2 ("mostly good") is rather large?

Yes, it's a balancing act between being too coarse (0/1) and too fine (too many categories for a user to choose from). We could have introduced an "in-between" category (1.5), but - analogous to reviewing papers - we wanted to encourage the annotators to make an overall decision about acceptability (2+) of the output, rather than just say a less helpful, undirected "in-between" assessment.

### Q: How do you ensure the quality of crowd annotation?

To ensure the quality of our crowd worker annotation, here are some measures we took:

- (a) In the design of the task, we ask workers to follow the guidelines for explanation critiquing to identify significant flaws and rate the explanation quality first. This stage is designed to familiarize the workers with the task by trying it out themselves and to let them form their own judgment regarding the quality of the explanation. After which, we then ask them to rate the GPT-4 generated critique. This ensures that their judgment is not just swayed by the generated critique.
- (b) In the “rules” section of our instructions to crowd workers, we noted: *“Finally, note that if you work on this HIT, please answer with care: Some HITs will be checked by hand, and work may be rejected if there are too many errors (e.g., rating a poor critique with a high critique score of 3).”*
- (c) During data collection, we first launched a set of 100 questions (along with explanations and critiques) as a pilot trial and sampled the annotations to check that our procedure yielded reasonable responses before launching the full set.
- (d) Percentages of critiques rated as good or perfect by crowd workers (92%) are comparable to those in expert annotations (86%).

**Q: How subjective are the critique dimensions? How much do annotators agree?**

Indeed, there could be a degree of subjectivity in identifying what is the most significant flaw dimension. From our annotation data, quantitatively, we observe that in 91.9% of the cases, there is at least some overlap between the flaw dimensions identified by the annotators (i.e. at least 2 annotators agree on a dimension). In terms of giving an explanation score, in 89% of the instances, all annotator’s explanation score is within 1 of the mean explanation score for that instance.

**B Details on Crowdsourcing for Estimating Quality of Explanation Critiques**

The crowdsourcing instructions included a detailed description of an ideal critique, along with four detailed examples for the different critique quality scores. The workers were asked to first provide their own set of explanation flaw dimensions as well as an explanation score (0-5) before being presented with the model-generated critique to give a quality score (0-3).

Figure 7 shows a screenshot of the initial part of the instructions to the workers.

**Turk workers and pay rate:**

Our participants were recruited on the Amazon Mechanical Turk platform. The workers met minimum qualification in AMT: 95% approval rate. They were from US locations and rated at Amazon’s Masters Level. The workers were paid at a rate of  $\approx \$15/\text{hr}$ .

**C Details of Explanation Critiquing task**

**C.1 Details on flaw localization,  $f_{loc}$**

In identifying the most significant flaw, when possible, we encourage quote(s) in this statement to better localize the flaw. In the quote(s), “...” is used to leave out parts not relevant to the flaw, and to keep just the erroneous parts. The quote(s) should be from the explanation, except in the case of pointing out a misunderstanding of the question and/or answer options, where the quote should try to identify the part of the question or answer that was misunderstood.

Following the spirit of the types of Socratic questions (Paul and Elder, 2006), we focus on more significant flaws and ignore more trivial criticisms e.g., upper-case/lower-case issues, small typos, and spacing inconsistencies.

**D Details of DS Critique Bank composition**

Table 7 shows details of the composition of each partition of DS Critique Bank. (Table 3 summarizes the main statistics of our DS Critique Bank.)

For science questions, the ARC dataset has both a Challenge Set and an Easy Set, where the Challenge Set contains only questions that are regarded as more difficult (answered incorrectly by retrieval-based and word co-occurrence algorithms).

DS Critique Bank partition	Questions from QA tasks	Explanations from student models	Critique instances	Human annotations
DSCB-train-silver	3240 questions: 810 ARC-Challenge-train + 810 ARC-Easy-Train + 270*6 Rainbow-Train	Each question with ONE of 4 student models * 3 explanation styles	3240 Q+A+E+Crit from GPT-4	NA
DSCB-train-crowd-anno	270 questions: 100 ARC-Challenge-Train + 50 ARC-Easy-Train + 20*6 Rainbow-Dev	Each question with 4 student models * 3 explanation styles	3240 Q+A+E+Crit from GPT-4	crowdsourced (all)
DSCB-train-expert	31 questions: 18 ARC-Challenge-Dev + 13 Rainbow-Dev	Sampling of 4 student models and 3 explanation styles	198 manually edited based on GPT-4 output	expert
DSCB-dev-crowd-anno	550 questions: 100 ARC-Challenge-Test + 50 ARC-Easy-Test + 50*6 Rainbow-Dev (no overlap with earlier usage) + 50 OBQA-Test + 50 CSQA-Dev	Each question with 4 student models * 3 explanation styles	6600 Q+A+E+Crit from critiquers: GPT-4 DS-7B DS-13B	crowdsourced (subset)

Table 7: Summary of data composition in DS Critique Bank.

The **DSCB-train-crowd-anno** partition was purposefully skewed towards challenging questions, by sampling such that the Llama-7B-chat student model had low 30% accuracy (vs the natural 60% accuracy). We favor sampling from Rainbow-Dev because the dev and test partitions of datasets like CosmosQA and Social IQa are designed to be of higher quality. CosmosQA sampled instances from workers who excelled in the generation task for their dev and test, while Social IQa’s dev and test were validated by workers a second time. In this partition, each critique is rated by 1 worker.

In **DSCB-dev-crowd-anno**, a large fraction (57%) of student model explanations receive a “None” critique from all 3 critique models, and this is less interesting to obtain annotations for. Therefore, in the subset to annotate (DSCB-dev-crowd-anno, subset), we keep only two such instances for each dataset, while randomly sampling instances for the other cases making sure there is at least one instance for every flaw dimension found by the GPT-4 critique model. In this partition, we obtain annotations from 3 workers for each critique and compute the rounded average of their ratings for our analysis.

In total, 26 different crowdworkers contributed to the annotations in DSCB-train-crowd-anno and 22 to annotations in DSCB-dev-crowd-anno. The specific anonymized worker ids have also been included in the released data, giving more details on how the annotations were distributed across workers.

## E Explanation Critiquing as a way to go beyond accuracy

### E.1 Examining quality of reasoning chain beyond accuracy

In Figure 8 we show how human annotated explanation scores vary for correct and incorrect answers across the four student models.

### E.2 Using explanation flaw dimensions to offer interpretability to explanation scores

Table 8 shows the explanation flaw dimensions across all student models and tasks. We see that even when explanation scores are similar and generally correlate with accuracy, the fine-grained dimensions indicate differences in models’ capabilities and weaknesses.

In Section 7.1 we already looked at the case study comparing GPT-3.5 and Llama2-70B student models on the Science datasets. We see from Table 8 that the smaller Llama2-7B model shares many of the same patterns as its bigger counterpart but with a larger fraction of “misunderstanding” and “inconsistent answer”.

Another observation when comparing models’ errors in science versus commonsense reasoning is that the flaws in their reasoning chains tend to be due to “incorrect information” for science reasoning compared to “misunderstanding” of the question or answer options when doing commonsense reasoning.

Such insights allow us a deeper understanding of models’ capabilities and can be useful for further research to improve model reasoning capabilities in a more targeted manner.

















Student model	Dataset	#Qs	Acc	E <sub>SC</sub>	Dimensions acc=1	Dimensions acc=0	E <sub>SC</sub> acc=1	E <sub>SC</sub> acc=0
GPT-4	Science	336	0.96	4.88			4.99	2.36
GPT-3.5	Science	319	0.84	4.43			4.87	2.15
Llama2-70B	Science	321	0.81	4.18			4.72	1.9
Llama2-7B	Science	278	0.48	3.11			4.69	1.63
GPT-4	Commonsense	267	0.82	4.37			4.87	2.12
GPT-3.5	Commonsense	248	0.63	3.71			4.76	1.88
Llama2-70B	Commonsense	221	0.62	3.23			4.25	1.6
Llama2-7B	Commonsense	227	0.44	2.6			3.95	1.55

Table 8: Analyzing critiques rated as good by crowdworkers shows that the type of errors in the reasoning chain varies across models and also depends on the type of task dataset. Refer to the legend in Figure 3 for dimensions of flaws. Explanation scores as a summary metric do not capture such nuances and characteristics across models and datasets.

### E.3 Allowing for fine-grained analysis through error localization and suggestions

Table 9 illustrates more examples of localized flaws across student models and domains. More examples of general suggestions are in Table 5. Such fine-grained information effectively uncovers specific errors associated with models and offers insights for future work on model improvement.

### E.4 Digital Socrates DS-13B analysis on individual datasets

In Tables 11 and 12 we provide the accuracy, explanation scores and flaw dimension pie charts produced by the DS-13B critique models for each of the 10 datasets in the DS Critique Bank dev set.

Analyzing these results across all 4 student models, reveals a rich diversity of behavior. E.g., on Physical IQa the models all tend to have more incorrect information than in other RAINBOW datasets, while CosmosQA, especially for GPT-3.5, has a large fraction of incorrect reasoning. Analyzing the individual instances for each dominant explanation flaw will provide more insights into model weaknesses.

### E.5 Comparing explanation flaws as judged by different critique models

In Figure 9, we show the full comparison of explanation flaw pie charts based on critiques from GPT-4, DS-7B, and DS-13B. We observe that the DS-13B and GPT-4 models, in particular, reveal very similar trends in student models’ flaw dimensions, while the DS-7B is a little less diverse in its critique dimensions,



and more likely to accept flawed explanations as having no flaws as long as the student correctly answers the question.

## **F More critique examples**

More examples from DS-13B are in Figures 10 (incorrect reasoning, incorrect information) and 11 (misunderstanding, None), covering other common flaw dimensions. These further examples of explanation critiques from GPT-4 vs DS-13B illustrate that, like GPT-4, DS-13B also provides high quality explanation critiques across different dimensions. Table 10 illustrates how DS-13B still provides correct and accurate critiques when the smaller DS-7B model is producing invalid critiques, suggesting that the former is better performing.

## **G Prompts for generating different explanations by student models**

The three different prompts used for generating explanations by the student models can be found in Figures 12, 13, 14.

## **H Prompt for generating explanations critiques from critique models**

The prompt used for generating critiques by the critique models (GPT-4, DS-13B, and DS-7B) can be found in Figure 15.

## **I Training hyperparameters**

The fine-tuning of the Digital Socrates DS-7B and DS-13B critique models were done on NVIDIA RTX A6000 GPUs, using the open-instruct code base.<sup>13</sup> The base models were Llama2-7B-Chat and Llama2-13B-Chat respectively. Low-Rank Adaptation (LoRA) (Hu et al., 2021) was used for parameter-efficient fine-tuning. The hyperparameters included: batch size = 32, lora rank = 64, lora alpha = 16, lora dropout = 0.1, learning rate = 1e-4, lr schedule = linear, warmup ratio = 0.03, weight decay = 0, flash attention. For the 13B model we also used QLoRA (Dettmers et al., 2023). A total of around 100 GPU hours was used between fine-tuning and inference for generating the DS Critique Bank dev set critiques.

---

<sup>13</sup><https://github.com/allenai/open-instruct>

When Artificial Intelligence (AI) systems such as ChatGPT answer questions, they can be asked to provide an explanation along with their answer as way of externalizing their reasoning. This provides us a window into their reasoning capabilities.

We have some systems that try to

- **identify** and **categorize** any main flaw in the generated explanation, as well as
- provide both **general** and **specific** suggestions to address the flaw.

In this task, after reading the instructions and the four examples, you will be asked to

We're seeking your help to understand (1) how good some current AI systems' explanations are and (2) how well systems can perform the task of **critiquing** explanations from AI systems. Please carefully review the instructions and examples, then perform the following two tasks on the **ONE** given question and explanation:

**[Task 1] Try critiquing the explanation yourself** (identify applicable flaws in the explanation, rate the explanation)

**[Task 2] Rate the 3 systems' critiques of the explanation** (i.e., judge the critique quality).

## [IMPORTANT] Ideal critique we're looking for:

**Part1: identify** and **categorize** any significant flaw in the generated explanation.

The ideal critique of a given answer and explanation should be helpful, like one that would be given by a knowledgeable tutor. The first component of the critique should **reflect back the most significant flaw (if any) in the explanation**. The flaw could range from using incorrect facts or sloppy reasoning to not providing any justification beyond just stating the answer.

The flaw should be **categorized into one of these 8 dimensions**:

- **misunderstanding**: apparent misunderstanding of the original question or answer choices
- **lack\_justification**: just stating the final answer without any proper or informative justification
- **incorrect\_information**: stating incorrect fact(s) or knowledge
- **missing\_information**: missing a crucial fact, knowledge, or perspective that should be considered
- **incorrect\_reasoning**: an incorrect leap in the reasoning chain
- **incomplete\_reasoning**: the reasoning doesn't lead all the way to the answer
- **inconsistent\_answer**: the answer doesn't match the conclusion of the explanation
- **irrelevant**: using irrelevant or redundant fact(s), knowledge, or reasoning

The flaw should be formulated as a **standalone erroneous statement**, that can be understood without reference to the full question or explanation.

When possible, we would like to have quote(s) in the reflection statement, where in the quote(s), "..." is used to leave out parts not relevant to the error, to leave just the erroneous parts. **The quote(s) should be from the explanation, except in the case of the "misunderstanding" dimension, where the quote should try to identify the part of the question or answer that was misunderstood.**

If necessary, **there could be a brief context to the reflection statement** so it can be read standalone without reference to the full question or explanation. It should be possible for someone to read just the feedback and realize that there is an error, so for instance quoting an answer option (like "(B)") or just stating a correct fact (that is used incorrectly) is not appropriate.

**If an explanation is reasonable**, and provides some meaningful evidence supporting the answer, **there is no need for the critique to be overly picky and possibly misleading, and we'd like "None" as the feedback instead.** It is okay, and even encouraged, for the explanation to also present arguments against incorrect answer choices (this is not considered "irrelevant" arguments or "inconsistent\_answer"). Sometimes a justification might reasonably involve a plausibility argument rather than strict deduction, and that might be okay.

**Part2:** provide both **general** and **specific** suggestions to address the flaw

The second component of the critique should provide statements toward fixing this flaw in the explanation.

- First, is a **general statement** that addresses a likely misconception or confusion behind the flaw. This should be so it **can be reused independently of this particular question**.
- Next, is a more **specific statement to consider when revising the explanation**.

**Part3:** finally, the explanation is to be **graded on a scale from 0 to 5**

The scoring guidelines are as follows:

- **0: Very wrong explanation**
- **1: At least has something useful in the explanation that's relevant to the question**

Figure 7: Part of the instructions to Amazon Mechanical Turk workers.

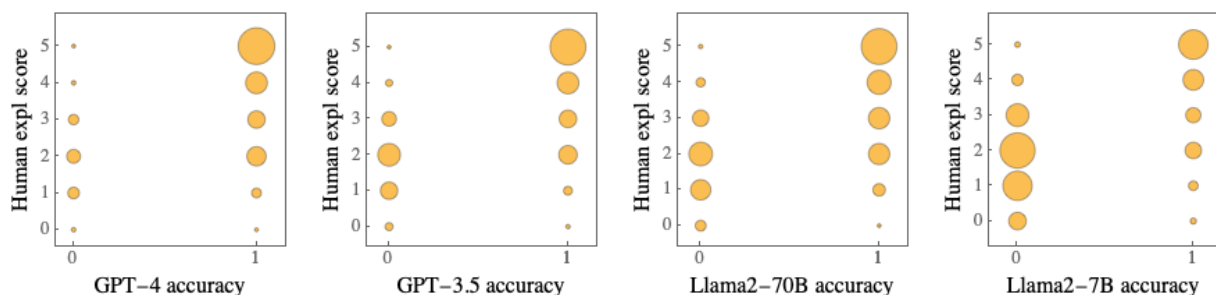


Figure 8: Across four student models, human-annotated explanation scores vary greatly within cases where models get the answer right (accuracy = 1) or wrong (accuracy = 0).

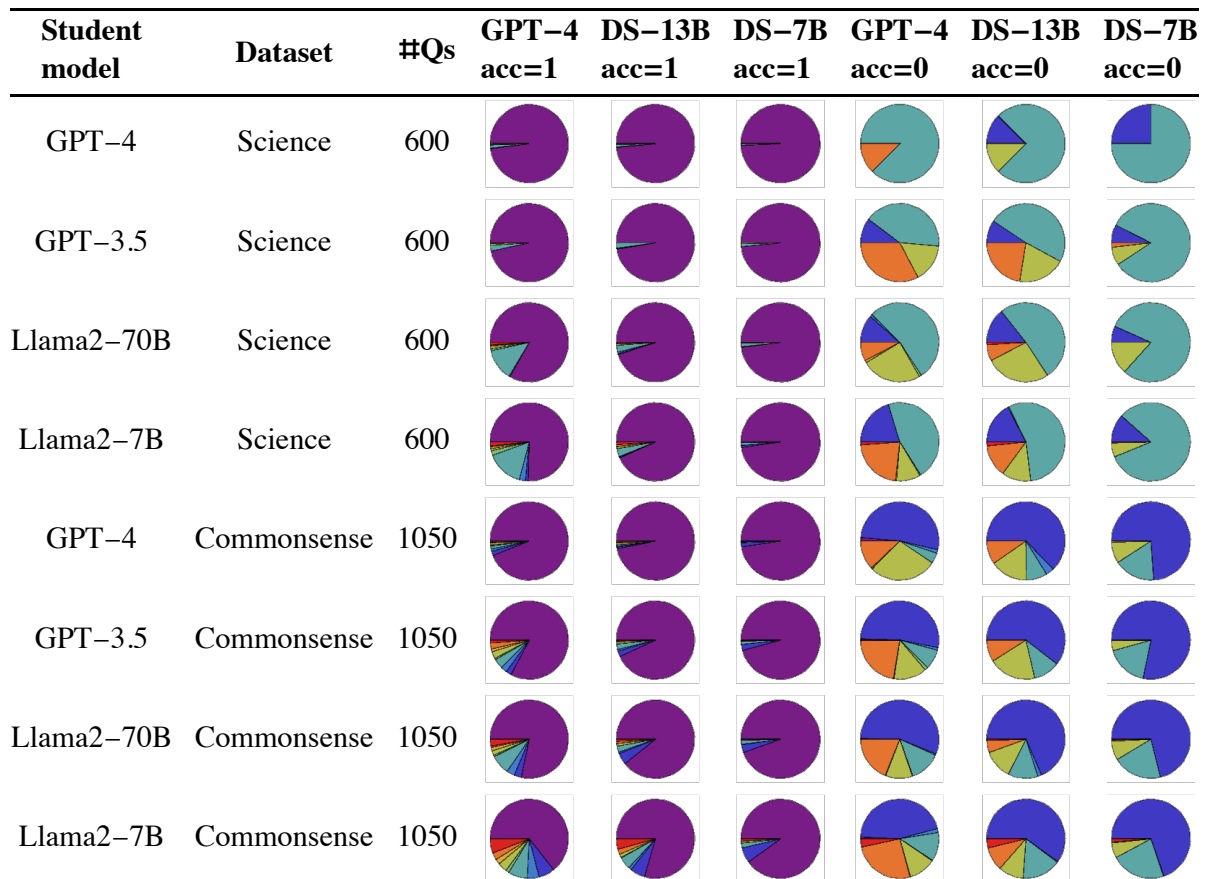


Figure 9: Comparison of insights on student models based on critiques from GPT-4, DS-7B, and DS-13B. On all data points in our dev partition (DSCB-dev-crowd-anno), the distribution of explanation flaws as identified by DS-13B is particularly close to that by GPT-4, especially in cases when student models got the questions wrong (acc=0).

Student model	Dataset	Dimension (acc=0) with frequency and example
GPT-4	Science	<b>incorrect information</b> (0.88): "The Sun, which is a sulfur burning heavenly body." <b>inconsistent answer</b> (0.12): "Therefore, the most likely outcome is a decrease in the population of blue flowers."
GPT-3.5	Science	<b>incorrect information</b> (0.41): "Option (D) is incorrect because it incorrectly states that birds must eat food to maintain a constant temperature, when in fact they can regulate their body temperature internally." <b>inconsistent answer</b> (0.33): "Given answer: (C)" while the explanation supports answer (D). <b>incorrect reasoning</b> (0.16): "Longer spines caused a genetic change in the fish." <b>misunderstanding</b> (0.1): "Thermometers indicate levels of mercury in the blood."
Llama2-70B	Science	<b>incorrect information</b> (0.53): "Ferromagnetic (C) materials are not typically used for paper clips." <b>incorrect reasoning</b> (0.25): "A decrease in the length of the ramp would result in a shorter distance over which the force must be applied, thereby requiring less force." <b>misunderstanding</b> (0.11): "Mutualism is a relationship in which both organisms benefit." <b>inconsistent answer</b> (0.08): "Given answer: (B)" while the explanation correctly describes clear and sunny weather, it incorrectly chooses option B as the correct answer. <b>incomplete reasoning</b> (0.01):
Llama2-7B	Science	<b>incorrect information</b> (0.46): "Longer spines caused a genetic change in the fish." <b>inconsistent answer</b> (0.22): "If the load is reduced by half, the force required to move the cart remains the same, so the cart will move faster (B)." <b>misunderstanding</b> (0.2): "Water changing from liquid to solid is described by option (C) snow melting into a puddle." <b>incorrect reasoning</b> (0.1): "Many of these bacteria became extinct due to changes in the Earth's environment that led to an increase in oxygen levels." <b>irrelevant</b> (0.01): "A mammal is a warm-blooded animal that has hair, produces milk for its young, and has a specific type of nervous system."
GPT-4	Commonsense	<b>misunderstanding</b> (0.52): "Without additional information, we cannot determine the speaker's preference." <b>incorrect reasoning</b> (0.28): "The location where he preferred to play and where he sounded quieter must be the same." <b>inconsistent answer</b> (0.12): "Given answer: (C)" while the explanation suggests that Bailey would feel cautious of things around her (B). <b>incorrect information</b> (0.04): "Smoke can be seen in the air due to the light scattering off particles, but it does not color the air itself." <b>lack justification</b> (0.01): "However, the context does not provide specific information about Jan's preferences or intentions, so any of the options could be possible depending on Jan's personality and mood."
GPT-3.5	Commonsense	<b>misunderstanding</b> (0.53): "The logical continuation would be that a final spot free rinse (A) is taken from under the hood." <b>inconsistent answer</b> (0.22): "Therefore, 'mortify' (B) is the best answer choice that aligns with the context of the sentence." <b>incorrect reasoning</b> (0.14): "Jordan would feel 'over' the relationship, indicating that they are ready to move on." <b>incorrect information</b> (0.08): "The cooking method that is more suitable for poor weather conditions is smoking, as it can be done indoors." <b>missing information</b> (0.01): "The context does not mention any reason for Natasha to go back inside the villa in the evening."
Llama2-70B	Commonsense	<b>misunderstanding</b> (0.56): "Quinn needs to work before he can afford nice meals." <b>inconsistent answer</b> (0.19): "Building a special room with no windows and no tech type of lights can create a sleep-conductive environment." <b>incorrect information</b> (0.13): "A stainless steel bucket (A) cannot be used to milk a cow into because it is not designed for that purpose." <b>incorrect reasoning</b> (0.11): "The writer asked if she should bring a swimsuit because it's good to wear for playing Scattergories Junior." <b>lack justification</b> (0.01): "The correct answer will describe the steps to open a bolt using a tool."
Llama2-7B	Commonsense	<b>misunderstanding</b> (0.45): "The sentence implies that Katrina had very short hair." <b>inconsistent answer</b> (0.26): "The context mentions that he was speaking at a mile a minute, which could cause him to not make sense." <b>incorrect information</b> (0.12): "A stainless steel bucket can be used to insulate against boiling water." <b>incorrect reasoning</b> (0.11): "This suggests that Natalie did not have a significant other at the time of the dance. Maria, on the other hand, did have a significant other, as she was able to leave the dance to get food." <b>irrelevant</b> (0.03): "To keep a plant healthy, it is important to provide it with adequate water, nutrients, and light."

Table 9: Examples of localized flaws sampled from human-verified critiques in DSCB-train-crowd-anno. The *explanation critiquing* task pinpoints the specific mistakes that models make under each dimension.



**Question:** A student is pushing a 20-kilogram box up a ramp. Which change will require the student to use less force to push the box? (A) increasing the mass of the box (B) decreasing the length of the ramp (C) decreasing the steepness of the ramp (D) increasing the friction on the surface of the box

**Explanation (Llama2-70B):** 1) The force required to push a box up a ramp is determined by the weight of the box and the angle of the ramp.

2) A decrease in the length of the ramp would result in a shorter distance over which the force must be applied, thereby requiring less force.

**Answer (Llama2-70B):** (B) [incorrect, should be (C)]

Critique GPT-4	Critique DS-13B
<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): "A decrease in the length of the ramp would result in a shorter distance over which the force must be applied, thereby requiring less force."</li> <li>* Dimension: incorrect_reasoning</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: The force required to move an object up a ramp is not directly related to the length of the ramp, but rather to the steepness of the ramp and the weight of the object. Shortening the ramp does not decrease the force needed, it just decreases the distance over which the force is applied.</li> <li>* Specific: In this case, the correct answer is (C) decreasing the steepness of the ramp. This would decrease the vertical component of the force required to push the box up the ramp, thereby requiring less force overall.</li> </ul> <p>Explanation score: 2</p>	<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): "A decrease in the length of the ramp would result in a shorter distance over which the force must be applied, thereby requiring less force."</li> <li>* Dimension: incorrect_reasoning</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: Remember that the force required to push an object up a ramp is determined by the weight of the object and the angle of the ramp, not the length of the ramp.</li> <li>* Specific: In this case, the force required to push the box up the ramp is determined by the weight of the box and the angle of the ramp, not the length of the ramp. Therefore, a decrease in the length of the ramp would not necessarily require less force.</li> </ul> <p>Explanation score: 2</p>

**Question:** When Dennis operates his lawnmower, he notices the engine makes a lot of noise. He also notices that the engine gets very hot. Which best describes the heat and noise generated from the lawnmower? (A) a change in phase (B) thermal expansion (C) an increase in entropy (D) mechanical advantage

**Explanation (GPT-3.5):** 1) The question states that the lawnmower engine makes a lot of noise.

2) The question states that the lawnmower engine gets very hot.

3) Noise and heat are both forms of energy.

4) The noise and heat generated from the lawnmower are a result of the conversion of energy from the fuel to mechanical energy.

**Answer (GPT-3.5):** (D) [incorrect]

Critique GPT-4	Critique DS-13B
<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): "The noise and heat generated from the lawnmower are a result of the conversion of energy from the fuel to mechanical energy."</li> <li>* Dimension: incorrect_information</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: It's important to understand that while noise and heat are indeed forms of energy, they are often considered waste or unwanted byproducts in many mechanical systems, not the desired output. This is because they represent energy that is not being efficiently converted into the intended form (like mechanical work), and instead is being dissipated in less useful ways.</li> <li>* Specific: In the context of this question, the noise and heat generated by the lawnmower engine are not the result of the conversion of energy to mechanical advantage, but rather represent an increase in entropy, which is a measure of disorder or randomness in a system.</li> </ul> <p>Explanation score: 2</p>	<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): "The noise and heat generated from the lawnmower are a result of the conversion of energy from the fuel to mechanical energy."</li> <li>* Dimension: incorrect_information</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: It's important to understand the difference between the different types of energy. Mechanical energy is the energy of motion, while thermal energy is the energy of heat.</li> <li>* Specific: In the case of the lawnmower, the noise and heat are not a result of the conversion of energy from the fuel to mechanical energy. The noise is a result of the vibration of the engine, while the heat is a result of the friction and combustion of the fuel.</li> </ul> <p>Explanation score: 2</p>

Figure 10: Further examples of explanation critiques from GPT-4 and DS-13B, covering the “incorrect reasoning” and “incorrect information” dimensions.

**Question:** Fill in the blank: The tip of James pencil was breaking while writing on the paper sheet. The \_\_\_\_ is weak. (A) pencil (B) paper

**Explanation (Llama2-70B):** The word "tip" refers to the pointed end of a pencil. The phrase "breaking while writing on the paper" implies that the pencil was being used to write on the paper. Therefore, the correct answer is (B) paper, as the sentence implies that the paper is weak.

**Answer (Llama2-70B):** (B) [incorrect, should be (A)]

Critique GPT-4	Critique DS-13B
<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"><li>* Main flaw (standalone statement): "Therefore, the correct answer is (B) paper, as the sentence implies that the paper is weak."</li><li>* Dimension: incorrect_reasoning</li></ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"><li>* General: Be careful not to misinterpret the cause and effect in a sentence. Just because the pencil tip is breaking while writing on the paper, it doesn't necessarily mean that the paper is weak. It could be that the pencil is weak, which is why its tip is breaking.</li><li>* Specific: In this case, the sentence "The tip of James pencil was breaking while writing on the paper sheet" suggests that the pencil's tip is weak, not the paper. The paper is simply the surface on which the pencil is being used.</li></ul> <p>Explanation score: 1</p>	<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"><li>* Main flaw (standalone statement): "The correct answer is (B) paper, as the sentence implies that the paper is weak."</li><li>* Dimension: misunderstanding</li></ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"><li>* General: Always ensure you understand the question and the context before making a conclusion. Misinterpretation can lead to incorrect answers.</li><li>* Specific: The sentence "The tip of James pencil was breaking while writing on the paper" implies that the pencil is weak, not the paper. The pencil is the one that is breaking, which suggests it is weak.</li></ul> <p>Explanation score: 1</p>

**Question:** Identify best explanation connecting the beginning and ending. Beginning: Karen didn't have a boyfriend for Valentine's Day. Ending: Her brother's thoughtfulness made her Valentine's Day special! (A) Karen decided to buy her brother a Valentine's gift. (B) Karens little brother handed her a box of chocolates.

**Explanation (GPT-3.5):** 1) The beginning states that Karen didn't have a boyfriend for Valentine's Day.

2) The ending mentions her brother's thoughtfulness made her Valentine's Day special.

3) The connection between the beginning and the ending should involve an action by Karen's brother that made her Valentine's Day special.

4) Giving a box of chocolates is a thoughtful gesture typically associated with Valentine's Day.

**Answer (GPT-3.5):** (B) [correct]

Critique GPT-4	Critique DS-13B
<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"><li>* Main flaw (standalone statement): None</li><li>* Dimension: None</li></ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"><li>* General: None</li><li>* Specific: None</li></ul> <p>Explanation score: 5</p>	<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"><li>* Main flaw (standalone statement): None</li><li>* Dimension: None</li></ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"><li>* General: None</li><li>* Specific: None</li></ul> <p>Explanation score: 5</p>

Figure 11: Further examples of explanation critiques from GPT-4 and DS-13B, covering the "misunderstanding" and "None" dimensions from DS-13B.

You are an experienced teacher who can answer the multiple-choice question by explaining why the single selected answer is most correct (and if relevant, why the other answers are incorrect). Produce a concise step-by-step explanation, at most 3 sentences. At the end clearly state the selected answer option, e.g., "Answer: (B)". Use the format:

Explanation: ...

Answer: ...

Here is the question: [[QUESTION]]

Figure 12: Zero-shot explanation prompt.

**Question:** Which illustrates the correct order of relationships in organisms? (A) tissue -> organ system -> organ -> cell (B) organ -> cell -> organ system -> tissue (C) cell -> tissue -> organ -> organ system (D) organ system -> organ -> cell -> tissue  
**Explanation (Llama2-70B):** The correct order of relationships in organisms is from the smallest to the largest unit, and it starts with the cell, then tissue, organ, and finally organ system. Therefore, the correct answer is (C) cell -> tissue -> organ -> organ system.  
**Answer (Llama2-70B):** (C) [correct]

Critique DS-13B	Critique DS-7B
<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): None</li> <li>* Dimension: None</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: None</li> <li>* Specific: None</li> </ul> <p>Explanation score: 5</p>	<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): "The correct order of relationships in organisms is from the smallest to the largest unit, and it starts with the cell, then tissue, organ, and finally organ system."</li> <li>* Dimension: incorrect_information</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: It's important to understand the hierarchy of biological structures and their relationships. The smallest unit is the cell, followed by tissue, organ, and organ system.</li> <li>* Specific: In the context of this question, the correct order of relationships is from the smallest to the largest unit, which is why option (C) is the correct answer. The order should be: cell, tissue, organ, organ system.</li> </ul> <p>Explanation score: 2</p>

**Question:** A bear cub learns to stay away from unknown bears because (A) they are much bigger than the cub (B) the other bears look like its mother (C) their mother teaches them to keep their distance (D) the unknown bears look harmless  
**Explanation (GPT-3.5):** The correct answer is (C) because the sentence states that the bear cub learns to stay away from unknown bears, implying that the mother bear teaches them to keep their distance. Option (A) is incorrect because the size of the unknown bears is not mentioned as the reason for staying away. Option (B) is incorrect because the sentence does not mention the appearance of the unknown bears resembling the cub's mother. Option (D) is incorrect because the sentence does not suggest that the unknown bears look harmless.  
**Answer (GPT-3.5):** (D) [incorrect]

Critique DS-13B	Critique DS-7B
<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): "The correct answer is (C) because the sentence states that the bear cub learns to stay away from unknown bears, implying that the mother bear teaches them to keep their distance."</li> <li>* Dimension: inconsistent_answer</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: Always ensure that your chosen answer aligns with your explanation. If your explanation supports a different answer, it's a sign that you've misunderstood the question or the answer choices.</li> <li>* Specific: In this case, your explanation correctly identifies that the mother bear teaches the cub to keep its distance, which aligns with option (C). However, you've chosen option (D) as your answer, which doesn't align with your explanation.</li> </ul> <p>Explanation score: 2</p>	<p>The explanation states or suggests the following:</p> <ul style="list-style-type: none"> <li>* Main flaw (standalone statement): "Option (D) is incorrect because the sentence does not suggest that the unknown bears look harmless."</li> <li>* Dimension: misunderstanding</li> </ul> <p>Consider these points for revising the explanation:</p> <ul style="list-style-type: none"> <li>* General: Always ensure you understand the question and the answer choices correctly. Misunderstanding can lead to incorrect conclusions.</li> <li>* Specific: In this case, the sentence does suggest that the unknown bears look harmless, which is the correct answer according to the answer sheet. The explanation should focus on why option (C) is the correct answer, not why option (D) is incorrect.</li> </ul> <p>Explanation score: 2</p>

Table 10: Failure examples where DS-7B is producing invalid critiques while DS-13B is producing valid ones.

Student model	Dataset	Dataset class	#Qs	Acc	E <sub>SC</sub>	Dimensions acc=1	Dimensions acc=0	E <sub>SC</sub> acc=1	E <sub>SC</sub> acc=0
GPT-4	ARC-Challenge	Science	300	0.98	4.93			4.98	1.6
GPT-3.5	ARC-Challenge	Science	300	0.84	4.42			4.93	1.82
Llama2-70B	ARC-Challenge	Science	300	0.82	4.33			4.87	1.85
Llama2-7B	ARC-Challenge	Science	300	0.58	3.48			4.8	1.67
GPT-4	ARC-Easy	Science	150	1.	4.92			4.92	
GPT-3.5	ARC-Easy	Science	150	0.86	4.5			4.93	1.86
Llama2-70B	ARC-Easy	Science	150	0.93	4.69			4.9	1.8
Llama2-7B	ARC-Easy	Science	150	0.7	3.95			4.89	1.78
GPT-4	OBQA	Science	150	0.98	4.91			4.97	2.
GPT-3.5	OBQA	Science	150	0.75	4.11			4.93	1.59
Llama2-70B	OBQA	Science	150	0.73	3.95			4.84	1.59
Llama2-7B	OBQA	Science	150	0.54	3.29			4.83	1.49
GPT-4	aNLI	Commonsense	150	0.79	4.25			4.93	1.75
GPT-3.5	aNLI	Commonsense	150	0.7	3.68			4.68	1.36
Llama2-70B	aNLI	Commonsense	150	0.66	3.57			4.67	1.45
Llama2-7B	aNLI	Commonsense	150	0.55	2.95			4.29	1.32
GPT-4	CosmosQA	Commonsense	150	0.79	4.23			4.88	1.71
GPT-3.5	CosmosQA	Commonsense	150	0.67	3.77			4.88	1.56
Llama2-70B	CosmosQA	Commonsense	150	0.65	3.67			4.86	1.51
Llama2-7B	CosmosQA	Commonsense	150	0.46	2.91			4.57	1.49

Table 11: Critique analysis from our trained Digital Socrates model DS-13B on the first 5 datasets in the DS Critique Bank dev set (out of 10 in total). Refer to the legend in Figure 3 for dimensions of flaws.



Student model	Dataset	Dataset class	#Qs	Acc	E <sub>SC</sub>	Dimensions acc=1	Dimensions acc=0	E <sub>SC</sub> acc=1	E <sub>SC</sub> acc=0
GPT-4	HellaSwag	Commonsense	150	0.9	4.66			4.97	1.87
GPT-3.5	HellaSwag	Commonsense	150	0.53	3.33			4.75	1.7
Llama2-70B	HellaSwag	Commonsense	150	0.54	3.13			4.56	1.46
Llama2-7B	HellaSwag	Commonsense	150	0.35	2.41			4.15	1.45
GPT-4	Physical IQa	Commonsense	150	0.92	4.51			4.79	1.33
GPT-3.5	Physical IQa	Commonsense	150	0.67	3.63			4.62	1.64
Llama2-70B	Physical IQa	Commonsense	150	0.73	3.83			4.67	1.52
Llama2-7B	Physical IQa	Commonsense	150	0.57	2.93			4.08	1.38
GPT-4	Social IQa	Commonsense	150	0.72	3.99			4.97	1.48
GPT-3.5	Social IQa	Commonsense	150	0.69	3.79			4.88	1.38
Llama2-70B	Social IQa	Commonsense	150	0.59	3.29			4.68	1.31
Llama2-7B	Social IQa	Commonsense	150	0.47	2.79			4.56	1.2
GPT-4	WinoGrande	Commonsense	150	0.74	3.99			4.97	1.18
GPT-3.5	WinoGrande	Commonsense	150	0.59	3.37			4.93	1.1
Llama2-70B	WinoGrande	Commonsense	150	0.55	2.95			4.37	1.25
Llama2-7B	WinoGrande	Commonsense	150	0.47	2.62			4.26	1.19
GPT-4	CommonsenseQA	Commonsense	150	0.76	4.25			4.97	1.97
GPT-3.5	CommonsenseQA	Commonsense	150	0.71	3.95			4.93	1.57
Llama2-70B	CommonsenseQA	Commonsense	150	0.75	4.09			4.92	1.66
Llama2-7B	CommonsenseQA	Commonsense	150	0.55	3.31			4.83	1.47

Table 12: Critique analysis from our trained Digital Socrates model DS-13B on the last 5 datasets in the DS Critique Bank dev set (out of 10 in total). Refer to the legend in Figure 3 for dimensions of flaws.

---

Answer the following multiple-choice questions and explain why each answer choice is either correct or incorrect.

Question: Which body system's primary function is the continuation of the species?

(A) digestive (B) nervous (C) excretory (D) reproductive

Reasoning: The continuation of the species requires children to be born, which is the primary function of the reproductive system (D). The other systems are related to ongoing functioning of the body: The digestive system (A) converts food into useful energy and nutrients. The nervous system (B) controls the communication between the brain and body functions. The excretory system (C) removes excess materials for the body.

Answer: (D)

Question: Fill in the blank: Amy always raised their voice loudly when they were angry but Christine did not because \_\_\_\_ had a calmer personality. (A) Amy (B) Christine

Reasoning: The blank is referencing the person with the calmer personality, and the sentence states that Amy (A) always raised their voice, while Christine (B) did not, thus Christine is likely the calmer personality.

Answer: (B)

Question: A scientist cross-pollinates a tomato plant that is resistant to viruses with a tomato plant that produces large tomatoes. This is an example of

(A) natural selection. (B) asexual reproduction. (C) selective breeding. (D) genetic engineering.

Reasoning: The scientist is selectively breeding (C) plants with desired traits, virus resistance and large tomatoes. Since the scientist interferes, it's not natural selection (A). It is also not asexual reproduction (B) because pollination is used. Because the genes are not directly manipulated, it's not genetic engineering (D).

Answer: (C)

Question: Which of the following best explains how stems transport water to other parts of the plant?

(A) through a chemical called chlorophyll (B) by using photosynthesis (C) through a system of tubes (D) by converting water to food

Reasoning: Water is transported in the stem through hollow tubes (C) in the xylem. Chlorophyll (A) and photosynthesis (B) is related to the separate process of converting water to food (D), rather than transportation.

Answer: (C)

Question: Context: Alex celebrated a birthday with Carson and they both had a fun time. Question: What will happen to Alex? (A) cry hysterically (B) have some food (C) buy a horse

Reasoning: The answer food (B) is suggested by the birthday setting, while nothing in the context is suggesting crying hysterically (A) or buying a horse (C).

Answer: (B)

Question: An engineer can choose from four different materials to make a heat-resistant handle for a stirring spoon. Which material is least likely to get hot when the spoon is placed in boiling water?

(A) Ceramic (B) Wood (C) Glass (D) Iron

Reasoning: The heat conductivity of the material will affect how hot the spoon gets when placed in boiling water. The heat conductivity of wood (B) is substantially lower than that of ceramic (A), glass (C), and iron (D).

Answer: (B)

Here is the next question, only answer this question and follow the pattern in the examples, starting with "Reasoning":

Question: [[QUESTION]]

---

Figure 13: Few-shot explanation prompt.

---

Answer the following multiple-choice question (each with 2 or more choices) by writing down a few verifiable, salient background statements which combine to explain the correct answer. First, here are some examples:

Question: Which of the following best explains how stems transport water to other parts of the plant? (A) through a chemical called chlorophyll (B) by using photosynthesis (C) through a system of tubes (D) by converting water to food

Reasoning:

1) Water is transported in plant stems through hollow tubes in the xylem.

Answer: (C) [1]

Question: Fill in the blank: Amy always raised their voice loudly when they were angry but Christine did not because \_\_\_\_ had a calmer personality. (A) Amy (B) Christine

Reasoning:

1) The sentence states that Amy always raised their voice loudly when they were angry.

2) The sentence implies that Christine did not raise their voice loudly when they were angry.

3) The person with the calmer personality is the one who did not raise their voice loudly, namely Christine. [1,2]

Answer: (B) [3]

Question: An engineer can choose from four different materials to make a heat-resistant handle for a stirring spoon. Which material is least likely to get hot when the spoon is placed in boiling water? (A) Ceramic (B) Wood (C) Glass (D) Iron

Reasoning:

1) The spoon gets hot when placed in boiling water.

2) A handle material with higher heat conductivity will get hotter.

3) Wood has substantially lower heat conductivity than ceramic, glass, and iron.

Answer: (B) [1,2,3]

Question: Which body system's primary function is the continuation of the species? (A) digestive (B) nervous (C) excretory (D) reproductive

Reasoning:

1) The continuation of the species requires offspring to be produced.

2) The primary function of the reproductive system is to produce offspring.

Answer: (D) [1,2]

Question: Context: Alex celebrated a birthday with Carson and they both had a fun time. Question: What will happen to Alex? (A) cry hysterically (B) have some food (C) buy a horse

Reasoning:

1) The context mentions a birthday celebration.

2) Birthday celebrations typically involve food.

3) There is no mention of any event that would cause Alex to cry hysterically or buy a horse.

Answer: (B) [1,2,3]

Question: A scientist cross-pollinates a tomato plant that is resistant to viruses with a tomato plant that produces large tomatoes. This is an example of (A) natural selection. (B) asexual reproduction. (C) selective breeding. (D) genetic engineering.

Reasoning:

1) Resistance to viruses is a desired trait.

2) Producing large tomatoes is a desired trait.

3) The scientist deliberately breeds plants with different, desired traits. [1,2]

4) Selective breeding involves choosing parents with particular traits to produce offspring with more desirable traits.

Answer: (C) [3,4]

Here is the next question, only answer this question and follow the pattern in the examples, starting with "Reasoning":

Question: [[QUESTION]]

---

Figure 14: Prompt for explanation with reasoning steps.

---

You are a knowledgeable tutor who gives helpful critique on a given answer and explanation to a question. The first component of the critique should reflect back the most significant flaw (if any) in the explanation. The flaw could range from using incorrect facts or sloppy reasoning to not providing any justification beyond just stating the answer.

The flaw should be categorized into one of these dimensions:

misunderstanding: apparent misunderstanding of the original question or answer choices  
lack\_justification: just stating the final answer without any proper or informative justification  
incorrect\_information: stating incorrect fact(s) or knowledge  
missing\_information: missing a crucial fact, knowledge, or perspective that should be considered  
incorrect\_reasoning: an incorrect leap in the reasoning chain  
incomplete\_reasoning: the reasoning doesn't lead all the way to the answer  
inconsistent\_answer: the answer doesn't match the conclusion of the explanation  
irrelevant: using irrelevant or redundant fact(s), knowledge, or reasoning

The flaw should be formulated as a standalone erroneous statement, that can be understood without reference to the full question or explanation.

When possible, use quote(s) in the reflection statement. In the quote(s), please use "..." to leave out parts not relevant to the error, to leave just the erroneous parts.

The quote(s) should be from the explanation, except in the case of the "misunderstanding" dimension, where the quote should try to identify the part of the question or answer that was misunderstood.

If necessary, try to add a brief context to the reflection statement so it can be read standalone without reference to the full question or explanation. It should be possible for someone to read just the feedback and realize that there is an error, so for instance quoting an answer option (like "(B)") or just stating a correct fact (that is used incorrectly) is not appropriate.

If an explanation is reasonable, and provides some meaningful evidence supporting the answer, there is no need to be overly picky and possibly misleading, just state "None" as the feedback instead. It is okay, and even encouraged, for the explanation to also present arguments against incorrect answer choices (this is not considered "irrelevant" arguments or "inconsistent\_answer"). Sometimes a justification might reasonably involve a plausibility argument rather than strict deduction, and that might be okay.

The second component of the critique should provide statements toward fixing this flaw in the explanation. First, try to provide a general statement that addresses a likely misconception or confusion behind the flaw. Formulate this so it can be reused independently of this particular question. Next, provide a more specific statement to consider when revising the explanation.

Finally, please grade the explanation on a scale from 0 to 5, with the scoring guidelines:

0: Very wrong explanation  
1: Has something useful in the explanation that's relevant to the question  
2: Makes some valid points, but is mostly incorrect  
3: Has many correct elements, but with significant flaws  
4: Mostly correct, with a minor flaw  
5: Completely correct

Please use the following format for the critique (use "None" if there is no feedback for a particular line):

The explanation states or suggests the following:

\* Main flaw (standalone statement): <flaw>  
\* Dimension: <dimension>

Consider these points for revising the explanation:

\* General: <reusable statement>  
\* Specific: <question-specific statement>

Explanation score: <score>

Here is the question with the given answer and explanation:

Question: [[QUESTION]]  
Correct answer (according to answer sheet): ([[ANSWERKEY]])  
Given answer: ([[PREDICTEDANSWER]])  
Given explanation: [[EXPLANATION]]

---

Figure 15: Prompt for generating critiques.