A Chain-of-Thought Is as Strong as Its Weakest Link: A Benchmark for Verifiers of Reasoning Chains

Alon Jacovi[1,2] **Yonatan Bitton**[1] **Bernd Bohnet**[3] **Jonathan Herzig**[1]
Or Honovich[1,4] **Michael Tseng**[3] **Michael Collins**[3] **Roee Aharoni**[1] **Mor Geva**[1,4]

[1]Google Research [2]Bar Ilan University [3]Google DeepMind [4]Tel Aviv University
alonjacovi@google.com

Abstract

Prompting language models to provide step-by-step answers (e.g., "Chain-of-Thought") is the prominent approach for complex reasoning tasks, where more accurate reasoning chains typically improve downstream task performance. Recent literature discusses automatic methods to verify reasoning steps to evaluate and improve their correctness. However, no fine-grained step-level datasets are available to enable thorough evaluation of such verification methods, hindering progress in this direction. We introduce REVEAL: *Reasoning Verification Evaluation*, a new dataset to benchmark automatic verifiers of complex Chain-of-Thought reasoning in open-domain question answering settings. REVEAL includes comprehensive labels for the relevance, attribution to evidence passages, and logical correctness of each reasoning step in a language model's answer, across a wide variety of datasets and state-of-the-art language models. Available at reveal-dataset.github.io.

1 Introduction

Complex reasoning tasks involve answering questions that require multiple steps of reasoning (Welbl et al., 2018; Talmor and Berant, 2018). Addressing these questions may require open-domain knowledge (Geva et al., 2021), mathematical reasoning (Cobbe et al., 2021; Hendrycks et al., 2021), logic (Dalvi et al., 2021), and so on. Reasoning chains—breaking the task into multiple steps explicitly—is useful for improving performance in such tasks, with LMs demonstrating better performance when encouraged to generate the reasoning chain behind their answer (Lampinen et al., 2022; Zelikman et al., 2022; Hu et al., 2023), commonly implemented via Chain-of-Thought (CoT) prompting (Wei et al., 2023, see Fig. 1 for an example).

Evaluation in such settings is traditionally limited to evaluating only whether the final answer is correct (Chowdhery et al., 2022; Wei et al., 2023;

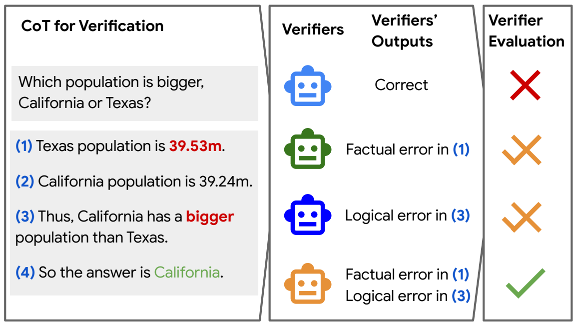


Figure 1: We collect REVEAL, an evaluation benchmark for the task of verifying reasoning chains in Chain-of-Thought format, which checks whether a reasoning chain is a correct justification to the final answer (importantly, the answer can be correct even if the reasoning is incorrect, as in the example above). The figure shows four verifiers (middle) verifying the correctness of a CoT (left). We use the dataset to benchmark multiple verifiers (right).

Wang et al., 2023). However, correct reasoning chains have been shown to be correlated with better final answers (Jung et al., 2022), with recent literature proposing automatic methods for verifying the quality of the reasoning chains themselves along various axes such as informativeness, relevance, factuality and logical correctness (Golovneva et al., 2023; Press et al., 2023; Opitz and Frank, 2021; Leiter et al., 2022). While such verification methods are a promising direction for improving reasoning in LLMs, it is not clear how to evaluate them due to the lack of high-quality, step-level annotated data, and collecting such data was shown to be difficult (in terms of reaching high inter-annotator agreement) and costly (Golovneva et al., 2023).

We present REVEAL (*Reasoning Verification Evaluation*), an evaluation benchmark for complex reasoning verifiers.[1] REVEAL covers a diverse set of reasoning skills, complexity levels, and knowl-

[1]At huggingface.co/datasets/google/reveal. We adopt practices by Jacovi et al. (2023a) against data contamination and request that any future redistribution or usage of the data respects the same constraints.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4615–4634
August 11-16, 2024 ©2024 Association for Computational Linguistics

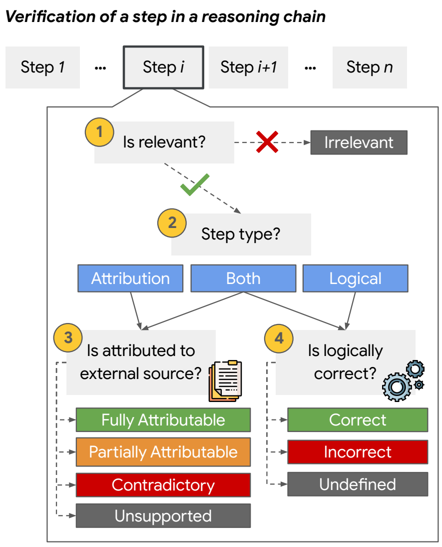


Figure 2: A flowchart of our protocol for verifying reasoning correctness step-by-step (§2).

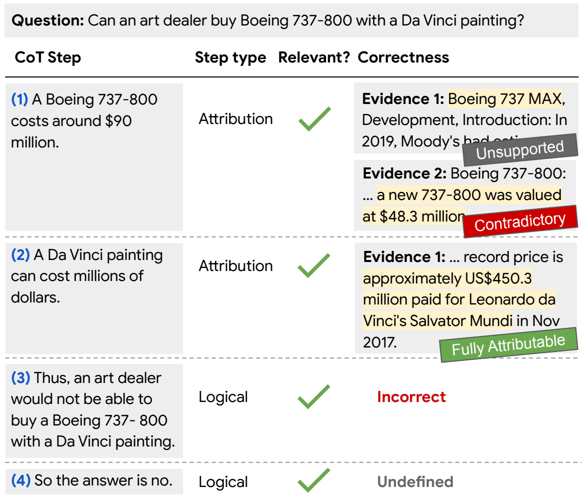


Figure 3: A REVEAL instance, with labels for step type, relevance, and correctness (attribution to a source or logical correctness from previous steps). Each label is accompanied by a high-quality free-text justification (shown in Fig. 4). We retrieve up to three evidence paragraphs, until a non-"unsupported" evidence is found. For Step (4) logical correctness is undefined, since it follows a logically incorrect step (Step 3).

edge domains. It contains 704 unique questions from 4 popular QA datasets, and 1,002 CoT answers generated by 3 language models, consisting of 3,360 CoT steps in total.

Each step is first labeled for relevance with respect to the final answer, and then whether the step is an attribution step (introduces factual knowledge which can be attributed to a source), a logical step (introduces logical inference from previous steps) or both. For attribution steps, we collect labels for correctness to retrieved Wikipedia paragraphs given as evidence (with full support, partial support, contradiction, or no-support as labels). For logical steps, we label for logical correctness. Each label includes free-text justifications written by the annotators. An illustrative instance from the dataset is shown in Figures 3 and 4. We split the dataset into REVEAL-Eval, the main evaluation benchmark containing high inter-annotator-agreement labels, and REVEAL-Open, a smaller set of interesting borderline cases with open labels due to low inter-annotator agreement. In §5 we describe the dataset, and report fine-grained analyses of non-attributable steps in REVEAL-Eval (i.e., evidence that supports or contradicts them was not found) and of disagreement categories in REVEAL-Open.

REVEAL supports versatile evaluation settings, for example: (1) Attribution steps, along with their evidence, can serve as a high-quality Natural Language Inference (NLI, Dagan et al., 2005; Bowman et al., 2015) benchmark in a setting of fact-checking LM outputs (Gao et al., 2023; Zhang et al., 2023); (2) CoT verifiers can be evaluated at the level of individual steps, or (3) at the level of full CoT answers; (4) Each label in the data contains five free-text justifications (one per annotator), which can accommodate research around the generation of explanations and justifications, or be used to understand nuance in borderline cases.

As we focus on the evaluation of step-level validation in complex reasoning, in §6 we report the performance of multiple up-to-date verification baselines, leveraging NLI classifiers, GPT-3 and PaLM 2, showing much room for improvement in current state-of-the-art solutions. In particular, verifiers struggle at classifying whether a step conveys correct logical inference from previous steps.

In summary, this work includes the following contributions: (I) A protocol for step-by-step verification of reasoning chains (§2); (II) An annotation schema to reliably execute the protocol with human annotators (§3); (III) A new benchmark dataset for evaluating automatic reasoning chain verifiers (§4, §5); (IV) Detailed analyses of challenges in retrieving evidence to knowledge claims in reasoning and documentation of disagreements in the data (§5); (V) A study on the challenges for current verifiers (§6). These contributions advance the research on verification of reasoning chains and methods for correctly reasoning about complex questions.

2 Formalism for Verification of Reasoning Chains

In this section, we formalize the task of verifying reasoning chains for attribution and logical correctness and discuss its evaluation.

We consider a reasoning chain as a sequence of n steps $r = s_1, ..., s_n$, where s_i is a claim generated conditionally on the steps preceding it $s_1, ..., s_{i-1}$. We focus on reasoning chains that are generated to answer a given question q and have a CoT format, that is, where the last step s_n includes the final answer to q and every step is a standalone sentence.

Given vocabulary $\mathbb{Z}$, a verifier $\mathcal{V} : \mathbb{Z}^{|q|} \times \mathbb{Z}^{|r|} \rightarrow [0, 1]$ receives a question q and a reasoning chain r conditioned on q, and outputs a score for the correctness of r as an answer to q. "Correctness" can be defined in various factors, such as factuality, grammaticality, and coherence. We focus on two key factors—correctness with respect to world knowledge, i.e., that the answer is derived based on grounded facts, and logical correctness, where reasoning steps are inferred with logically correct inference. We consider the world knowledge that supports the claims in r as some evidence e, external to the reasoning chain. For our purposes, we separate the retrieval of e from the verification process. Therefore, we consider a verifier $\mathcal{V} : \mathbb{Z}^{|q|} \times \mathbb{Z}^{|r|} \times \mathbb{Z}^{|e|} \rightarrow [0, 1]$ that also receives as input evidence e to ground the correctness of r.

Full chain vs. step-level verification. At high-level, a verifier is a system which receives as input a question and a reasoning chain answer (in our case, a CoT answer), and outputs whether the *entire* reasoning chain is correct—this is the more prevalent approach (§7). A more fine-grained approach is to evaluate each reasoning step for correctness separately, where the reasoning chain is correct if every step is correct (Li et al., 2023). We adopt the step-level approach, as it provides the ability to evaluate full-chain verifiers, the ability to detect the exact point of failure in reasoning chains, and the ability to distinguish between different types of errors. Step-level detection can additionally help with detecting cases of "snowballing of hallucinations" from earlier steps (Zhang et al., 2023).

Correctness of an individual step. Fig. 2 details our methodology for defining the correctness of an individual step, which we expand on below.

(1) *Step relevance.* Each step is "`relevant`" or "`irrelevant`" to answer the question. Irrelevant steps do not invalidate the chain's correctness.

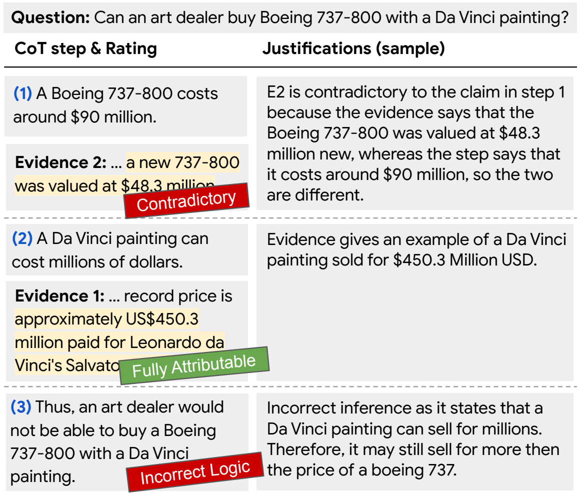


Figure 4: Example free-text justifications for the labels in REVEAL. There are 5 justifications for each label.

(2) *Step type.* The correctness of a step can be defined in three ways: Whether it is entailed from approved world knowledge ("`attribution step`"), entailed from previous steps ("`logical step`"), or both ("`both`"). We define attribution steps in language adopted from fact-verification (Konstantinovsky et al., 2018; Guo et al., 2022): If the step *"contains knowledge that should be verified against an external source."*

(3) *Step attribution to external source.* An attribution step is correct if it is "`fully attributable`" to a given source, meaning "strictly according to the given source, all of the information in the claim is correct." Steps are otherwise "`contradictory`" or "`partially attributable`" to the source, or "`unsupported`" by the source. This categorization mimics the Attribution to Identified Sources formalization (AIS, Rashkin et al., 2021).

(4) *Step logical correctness.* Each logical step is "`correct`" if it can be logically inferred from the previous steps, otherwise it is "`incorrect`". The correctness of logical steps that follow incorrect logical steps is undefined.

(3 & 4) *Hybrid steps.* When steps contain both world knowledge and logical inference, we assume that all external knowledge is fully attributable for the purpose of logical correctness, and that all logical inferences from previous steps are correct for the purpose of attribution correctness.

3 Annotation Schema

We introduce a labeling schema for reasoning chain verification, according to the formalization in §2.

The schema prioritizes annotation quality, ease of use and scalability. Such an implementation requires making some design choices, which we explain and justify below. The annotation interfaces and annotation questionnaire are available in §A.

We observe that the overall annotation process of a reasoning chain involves different types of information and requires various skills. For example, to determine if a reasoning step can be attributed to a given evidence paragraph, it is essential to read the paragraph in detail and judge its relation to the step without any additional context. In contrast, to validate if a reasoning step is logically correct, it is required to think critically whether some information is missing or incorrect in the overall reasoning process while leveraging the annotators' background knowledge. The two modes of thinking (logic and attribution) can interfere with each other, where focusing on one at a time is less cognitively-demanding.

To accommodate the complexity of the verification task, we therefore split it into two annotation tasks: (1) A full-chain task that considers the logical transitions between steps, while assuming that any specified facts are factual; (2) a step-wise task that checks if a given step is attributable to or contradicts some given external knowledge.

Note that *the two tasks are designed to be independent*, such that different annotators can annotate logical correctness independently from attribution, and vice versa. This allows for parallelization of the annotation process, reduces cognitive load, and provides a more robust set of labels that is less dependent on specific annotators. For example, some annotators were assigned to only one of the tasks, according to feedback during pilot phases.

Justifications. In addition to the verification labels, in both tasks we ask annotators to provide a free-text justification for their choices. This information helps to monitor the annotation process, provide additional valuable insight on the reasoning behind ratings—as it is possible to make the same decisions in these tasks for different reasons—and additionally serves as a method of interacting with the annotators to provide and receive feedback.

3.1 Task 1: Relevance, Type, and Logic

Given a complex reasoning question and a CoT answer, the first task involves annotating the chain step-by-step, labeling the following information for each step: *relevance*, *step type* and *logical correctness*, according to the classes described in §2.

The logical correctness of a step in the chain is evaluated with respect to the steps preceding it and the question. Importantly, this is done without considering the correctness of attribution steps that introduce external knowledge—i.e., all external knowledge is assumed to be correct in this task, as the focus is on the logical derivations. Consider for example the following reasoning chain:

> *Question:* Which population is bigger, California or Texas?
>
> *Answer:* Texas population is 39.53m.[(1)] California population is 39.24m.[(2)] Thus, Texas has a bigger population than California.[(3)] So the answer is Texas.[(4)]

When evaluating logical correctness, step 3 would be considered correct, even though the cited number for the population of Texas is incorrect.

3.2 Task 2: Relevance and Attribution

The second task considers the correctness of facts stated in the reasoning chain and therefore is applied only to steps labeled as attribution steps (in the first task). The task involves verifying every step against a set of evidence paragraphs, where the annotator should indicate for a given evidence paragraph if the step is fully attributable (supported), partially attributable, contradicts or is unsupported. The annotation of a step ends when either a fully-supporting or contradicting evidence is found, or the maximum number of paragraphs is labeled. In this work, we limit this number to three, as evidence support dramatically decreases with additional evidence (Fig. 5 right) and to reduce cost.

Relevance is annotated in this task as well, to account for more specific phrasing under the lens of attribution: An attribution step is marked relevant here if it introduces information that is helpful to answer the question. When writing justifications to their ratings, annotators were encouraged to quote specific parts of the CoT or evidence as needed.

4 Data Collection Process

In this section, we describe the full process for constructing REVEAL, using our annotation schema. The process is divided into two phases: Collecting and generating data for annotation (steps 1–2) and the annotation process (step 3).

Step 1: Reasoning Chain Generation. We use four open-domain complex-reasoning QA datasets to elicit model-generated reasoning chains:

1. StrategyQA (Geva et al., 2021): Yes/no questions that require a diverse set of reasoning skills and applying implicit knowledge.
2. MuSiQue (Trivedi et al., 2022): Multi-hop reasoning questions with free-text entity answers, generated based on paragraphs from Wikipedia. For our dataset, we consider the subset of 2-hop questions, as we found qualitatively that the 3-hop and 4-hop questions are often unnatural and hard to understand.
3. Sports Understanding (Srivastava et al., 2023): Yes/no questions that require reasoning about knowledge on sports players, leagues, and sports maneuvers.
4. Fermi (Kalyan et al., 2021): Estimation questions with numerical answers that require both knowledge and reasoning to answer. The questions are designed to have no clear gold answers—they are exercises of common-sense reasoning and numerical estimation, e.g., "How much water does a school use in a week?"

Examples from each source are provided in Tab. 5 (§A). This diverse set of tasks provides a variety of question types, knowledge and reasoning requirements, and answer formats (binary, numerical and free-text entities). The questions were sampled randomly from the evaluation sets of each dataset and evenly split across the four datasets.

We use three different LMs to generate CoT answers: Flan-PaLM-540B (Chung et al., 2022), GPT-3 (text-davinci-003, Brown et al., 2020) and Flan-UL2-20B (Tay et al., 2023). We prioritized a variety in the models, with two large-size high-performing models with different pretraining data, and one smaller-size, weaker model, in order to gather a variety of CoT answers. CoT prompt demonstrations were written according to standard practices (Wei et al., 2023), with in-domain examples for each dataset taken from the dataset's training set, and designed to be simple and informative. A subset of the questions from each dataset (one-fourth) were answered by all three models with three separate CoT answers, to serve as a base for analyses that require multiple answers per question, while the rest were answered by each model once in equal proportions (one-fourth each). Our goal in this methodology was to maximize the flexibility of the dataset and potential for analyses and evaluations, for as many LMs and verifiers as possible. Specifically, we aimed to collect a sufficient variety of answers, *both correct and incorrect*, from a practical and realistic distribution.

	Reveal-Eval	Reveal-Open
Questions	704	205
CoT answers	1002	224
CoT steps	3360	847
Avg. steps per CoT	3.4	5.1
Attribution steps	1979	485
Step-evidence pairs	3502	745
Avg. evidence length (words)	103	103
Logic steps	1250	306
Fully attributable step-evidence pairs	864	–
Logically correct steps	1063	–
Fully correct CoT answers	200	–

Table 1: Quantities for various properties of Reveal.

Step 2: Evidence Retrieval. For attribution verification, we use Wikipedia as an external knowledge source and retrieve three paragraphs for each attribution step. StrategyQA, MuSiQue, and Sports Understanding are well-supported by Wikipedia, while Fermi is explicitly designed to be difficult to support, giving a variety of cases. To promote the retrieval of supporting paragraphs, we mix dense retrieval and lexical-based retrieval, fetching two paragraphs with GTR (Ni et al., 2021) and one with BM25 (Zaragoza et al., 2004). In addition, retrieval is done using decontextualized versions of the CoT steps, generated by the decontextualization model of Choi et al. (2021) which replaces co-referenced pronouns with explicit entity mentions.

Step 3: Annotation. We use a pool of 13 English-speaking annotators, and collect 5 annotations per question and its corresponding CoT answer, for each of our two tasks (see §3). A small portion of the labels (approx. 5%) were annotated by the authors of this work (3 annotations per sub-task), to fill any gaps in examples that were not fully annotated for technical reasons.

Justifications and Quality Validation. Each annotation in Reveal is accompanied by a free-text justification written by its annotator. Those justifications serve two goals: they allow us to easily monitor the annotation process by leaning into the annotators thought process and provide them with meaningful feedback (Nangia et al., 2021), and serve as a valuable resource on their own which we leave to be explored in future work. We maintained a high level of quality for these justifications, such that they do not collapse to templated answers or provide insufficient information. Examples of these justifications are given in Fig. 4.

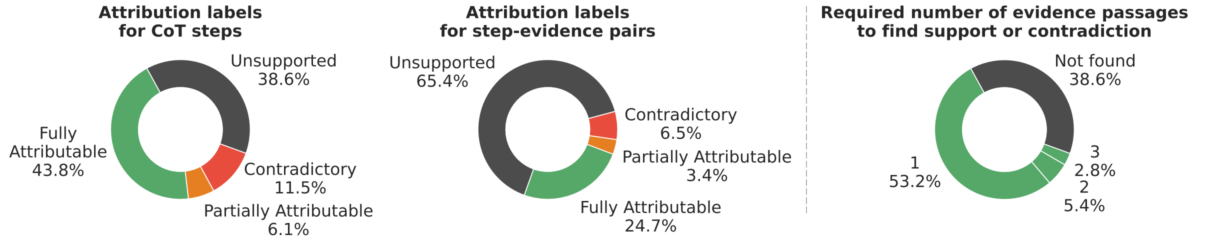


Figure 5: Statistics for the attribution task in REVEAL-Eval (§5). Left: Label distributions for each step or step-evidence pair. Right: Statistics for the number of retrieved evidence passages for each attribution step until a supporting or contradicting evidence was found (up to 3).

In addition, we performed three pilot rounds for annotator selection, annotator training, and task improvement, all to maintain a high quality bar for the collected dataset (the pilot annotations were discarded from the final dataset). Finally, we split the dataset into two subsets, REVEAL-Eval and REVEAL-Open, where the latter contains low-confidence labels that received lower inter-annotator agreement (see §5).

5 REVEAL

We ran the data collection protocol described in the previous section to obtain our dataset, REVEAL. In terms of inter-annotator agreement, we report a Krippendorf's α of 0.49 for attribution steps and 0.46 for logical steps.

We split REVEAL into two subsets, as we observe that some reasoning chains are very challenging to annotate because of ambiguity or other factors (analyzed in §5.2). Any CoT answer with at least one step that has low inter-annotator agreement is treated as an indecisive case, which applies to 18% of the CoT answers. A step is considered to have "low inter-annotator agreement" if less than three annotators agree on any single label for attribution or logical correctness (for example, an attribution step labeled as "partially attributable" by two annotators, "fully attributable" by one annotator, and "contradictory" by two annotators). We release these examples separately in a data subset called **REVEAL-Open**, as they are valuable instances of difficult or borderline cases, and the rest of the examples in a subset called **REVEAL-Eval**. Since the annotations in REVEAL-Open are indecisive, in §6 we evaluate only on REVEAL-Eval.

Tab. 1 details statistics on the two data splits, showing that REVEAL-Eval has several hundreds to thousands of examples for each setting to support reliable evaluation. Instances are approximately distributed uniformly across source datasets and

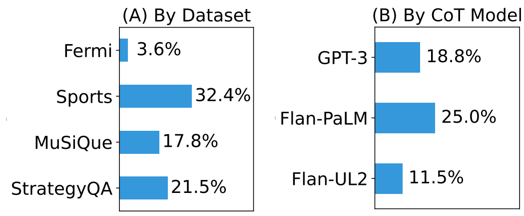


Figure 6: Comparison of full CoT correctness across subsets in REVEAL-Eval, meaning that *all steps* in the CoT are fully attributable and logically correct. (B) only includes questions answered by all models, where all answer CoTs are in REVEAL-Eval.

LM generators. We release all five annotations for each label in the dataset, including anonymized annotator identifiers for each label, to support future annotation methodology research and reproducibility (Sandri et al., 2023).

Considering the label distributions, 98.6% of steps are relevant to answer the question, and 87.5% of logic steps are logically correct. Fig. 5 (left) shows label distributions for attribution steps and step–evidence pairs (note that partial attribution is considered as "not enough info" in standard NLI formalization). While 43.8% of attribution steps were fully attributable, this is still relatively low, which may point at attribution as a main obstacle in reasoning. However, we do note that unsupported steps are not necessarily incorrect, as lack of support may stem from wrong claims, but also from issues in retrieving relevant supporting evidence. We analyze this further in §5.1.

Fig. 5 (right) shows the distribution for the required number of evidence paragraphs for finding support or contradiction for each step: In the majority of cases where support was found, it was already present in the first retrieved paragraph.

Fig. 6 shows distributions for full chain-level attribution and correctness for each subset of the data. In terms of specific types of errors in the CoTs: 77.3% of CoTs in REVEAL-Eval have a step which is not fully attributable, while 18.5% have a step which is logically incorrect.

5.1 Analysis of Unsupported Claims

A large fraction of the labeled steps was found unsupported by the retrieved evidence (38.6%, Fig. 5). To understand this, we conduct a qualitative analysis of 40 unsupported steps (10 from each source dataset, randomly sampled). For each step, we first verify whether the step is factually correct. For correct steps, we analyzed whether the issue was due to irrelevant retrieved evidence, or due to other reasons. See §A for full details.

Of the 40 unsupported steps, 19 were indeed not factual. Of the remaining 21: In 13 steps the retrieved evidence was relevant, but additional reasoning and world knowledge was required to verify them (e.g., which teams are part of the spanish "La-Liga", or that "ACM" stands for "Association for Computing Machinery"). In 6 steps the retrieved evidence was irrelevant. In 1 step there were two claims within the given step, but only one was supported by the evidence, and in 1 step the evidence was irrelevant due to insufficient decontextualization ("Brown" ref. "Jaylen Brown").

From this analysis, we estimate that roughly half of the unsupported cases stem from using imperfect retrieval. Additional challenges such as imperfect decontextualization and requiring to reason over multiple evidence passages can also slightly contribute to failing to support factually correct claims.

5.2 Analysis of REVEAL-Open

REVEAL-Open consists of questions paired with answers in which at least one of the steps has low inter-annotator agreement, defined as a case where no more than two (out of five) annotators agree on a single label for that step. Tab. 2 summarizes our analysis of the low-agreement cases. We inspected each of these steps (including the underlying data, as well as annotators' individual labels and written feedback) and report a qualitative description of the complications that may have contributed to the low agreement. On the basis of these qualitative notes, we devise 13 broader categories and manually assign one or more categories to each step.

The most frequent complication categories for attribution steps are world knowledge or general inference, rating category definition/criterion, and specialized knowledge and insufficient nuance/hedging tied. For logical steps, the most frequent categories are calculation or unit issue, inconsistent/unclear reference or standard, and invalid

Complexity Category	Attribution steps (N=199)	Logic steps (N=52)
Approximating and hedging	1.51%	–
Averages and ranges	10.05%	–
Calculation or unit issue	–	42.31%
Specialized knowledge	13.07%	7.69%
Formatting issue	3.02%	–
Unclear reference	12.56%	30.77%
Inference across claims in evidence	1.51%	–
Invalid inference in previous step	0.50%	28.85%
Insufficient hedging	13.07%	–
Rating category definition/criterion	23.62%	–
Relevance dispute	–	23.08%
Temporal inconsistency	9.55%	1.92%
World knowledge	25.63%	15.38%

Table 2: Percentage of steps per complexity category in REVEAL-Open. A step can match multiple categories. For more details see §A.

inference in a previous step.

We conjecture that improvements to the annotation instructions could help to minimize some of the categories, despite our efforts to make revisions during the pilot phases to achieve this effect. In particular, additional guidance can assist decisions on when and how to apply world knowledge, and clarify distinction between attribution labels. Recruiting annotators with appropriate domain knowledge for the main subjects of a given dataset (e.g., sports or sciences) could also reduce the number of disagreements related to both the content and the standards of comparison or equivalence.

6 Experiments

We use REVEAL to evaluate existing methods for CoT verification, on step-level verification (§6.1) and CoT-level verification (§6.2).

Verifiers. We use Flan-UL2 (20B), Flan-PaLM (540B), PaLM-2-L (Anil et al., 2023) and GPT-3 (`text-davinci-003`) for LM baselines in few-shot prompting settings (prompt templates detailed in §B). In addition, we use two specialized baselines: a T5-based model with 11B parameters trained on a mixture of NLI datasets (Honovich et al., 2022), and FacTool—a GPT-3-based fact-checking pipeline (Chern et al., 2023).

For the prompting-based baselines, we use the label generated by the LM as the predicted class (which always matched to one of the task labels in our experiments). The NLI classifier receives as input a premise and hypothesis and predicts a score between 0 and 1 that indicates if the hypothesis is entailed by the premise or not. FacTool returns a binary factuality classification label, and we use it

Baseline	Attribution 2-class	Attribution 3-class	Logic	Type
Flan-UL2-20B	65.2	50.4	59.4	27.3
Flan-PaLM-540B	85.1	66.0	68.6	51.2
PaLM-2-L	85.9	**70.7**	**77.6**	**64.1**
GPT-3	81.4	51.3	59.4	52.3
FacTool	71.1	–	–	–
t5-xxl-true	**88.4**	55.0	47.3	–
Class balance	76:24	70:24:6	80:20	59:40:1

Table 3: Macro-F1 for all step-level tasks in §6. Performance is measured across steps.

Baseline	Single decision	Pipeline
Flan-UL2-20B	41.5	54.4
Flan-PaLM-540B	39.4	58.1
PaLM-2-L	61.9	**76.4**
GPT-3	35.6	71.9

Table 4: Macro-F1 for the CoT correctness task (§6). This task involves *only* specifying whether a CoT is correct or not, regardless of the exact error. The pipeline variants use the step-level decisions (relevance, type, attribution and logic tasks) to check whether an incorrect step exists in the CoT.

only for the attribution task.

Evaluation. As few-shot evaluation is noisy (Perez et al., 2021; Jacovi et al., 2023b), we average the predictions of the LM baselines over 5 different 8-shot prompts, sampled in random order from 13 demonstrations for each sub-task, which we wrote for this purpose (the number of demonstrations is trimmed if it exceeds a given model's context length). All sets of demonstrations were class-balanced and balanced across source datasets and labels. Prompt templates are available in §B. For FacTool, we insert the appropriate evidence from the dataset, instead of allowing FacTool to retrieve evidence separately, for compatibility with the attribution labels. We release the prompt demonstrations alongside the dataset.

6.1 Step-level Verification

Given the unbalanced class distribution, we report macro-F1 performance in Tab. 3 (class F1 results are in §B). Results for the step relevance task (classifying if a step is relevant for answering the question) are omitted, as all the models collapsed to the majority baseline, leading to an F1 of near-0.

Step attribution. In this task, the model receives as input a decontextualized version of a CoT step and a Wikipedia evidence paragraph. There are two variants: *2-class*—classifying whether the step is entailed or not by the evidence; *3-class*—classifying between full entailment, contradiction, or not enough info. The best-performing model in 2-class was the T5-based classifier, which is much smaller than the other models although fine-tuned on large NLI data. 3-class performance follows scaling laws, which shows that detecting contradictions is a difficult problem.

Step logic. This task is to classify if a given step is logically entailed by the previous steps and the question. All prompting-based baselines were biased towards a "logically correct" classification, achieving high F1 on the "correct" class (>85%) and low F1 on the "incorrect" class (~33-47%). The T5 NLI baseline shows a bias towards the "incorrect" class, but performs significantly worse overall. This indicates that NLI fine-tuning with simple factual claims does not generalize to out-of-domain complex logical structures.

Step type. The task here is to classify whether a CoT step is an attribution step, logic step, or both, given the input question and its full answer. All models struggled with this task, achieving macro-F1 below 65%.

6.2 CoT-level Verification

This task involves verifying whether the CoT correctly justifies the answer or not—without necessarily specifying which step contains which type of error. Results are in Tab. 4. We report on two implementations of each LM baseline: (1) *Pipeline* implementations are decisions by combining the LM's decision on each step separately according to the step-level task predictions in §6.1. (2) *Single-decision* implementations are by prompting the LM to classify (in few-shot) whether a CoT is a correct answer to a question or not, given the question, CoT answer, and evidence paragraphs. The class balance for this task is 80:20 (incorrect CoTs being the majority class).

When observing macro-F1 performance, the pipeline variants significantly outperform single-decision in all cases. The gap appears to stem from a significantly higher F1 on the incorrect class (35-55% vs. 80-87%), while the correct class receives low F1 for both variants (30-56%). Overall, *all baselines struggled with the high-level task of CoT verification*, showing significant room for improvement on REVEAL and on CoT verification.

7 Related Work

While CoT prompting (Wei et al., 2023) has originally emerged as a simple prompting technique to increase performance of the "final" answer to the question, it has subsequently been interpreted as a part of the answer to be verified for correctness itself (Golovneva et al., 2023). As part of this narrative, various methods have been proposed to verify reasoning chains, either for LM evaluation (Welleck et al., 2022), improvement (Chen et al., 2023) or for training (Lightman et al., 2023). To our knowledge, no current work provides a fully labeled benchmark of reasoning chain correctness in information-seeking tasks.

Recent work on verification of reasoning chains primarily focuses on methods of verification and metrics: Golovneva et al. (2023) propose a suite of metrics for CoT quality, including knowledge attribution and logical correctness (both formalized as instances of NLI), grammaticality, informativeness, and so on. They provide a set of examples in math QA, closed-domain QA, and NLI, annotated in various quality metrics. Prasad et al. (2023) propose evaluating for informativeness and logical correctness in math QA and closed-domain QA settings. Ott et al. (2023) collect a variety of existing gold-reference human-written CoTs from existing datasets, and provide a web-interface annotation tool for annotating CoTs for errors in logical inference, factual knowledge, verbosity and reading comprehension. Lightman et al. (2023) provide a large dataset of math QA with step-level annotations for mathematical correctness of CoTs by one LM, primarily for training.

8 Conclusion

We design a methodology for human verification of reasoning chains, and employ it to annotate a dataset of CoT-format reasoning chains generated by three LMs in open-domain complex-reasoning QA tasks. The dataset, REVEAL, can be used to benchmark automatic verifiers of LM reasoning. Our work advances the research towards LMs that can provide correct and attributable reasoning behind their decisions. We find that CoTs generated by LMs are often not fully correct, and that automatic verifiers struggle to verify them appropriately. In particular, between attribution and logic types of correctness, *CoT-generating LMs* struggle more with attribution, but in contrast, *CoT verifiers* struggle more with verifying logical correctness.

9 Limitations

Evidence retrieval. Our work focuses on evaluation of verifiers that receive evidence as input—i.e., in the task of attribution to a given source, rather than the task of fact-checking where evidence retrieval is explicitly part of the task. As such, we can only evaluate verifiers that also operate on specific given evidence, rather than fact-chekers that perform evidence retrieval themselves. Additionally, for some of the knowledge claims in the dataset labeled as "unsupported", it is possible that there exists evidence that will support or contradict them, which our retriever did not surface. We note, however, that the goal of the dataset is to simply collect a sufficiently wide variety of cases on which to evaluate verifiers, rather than to evaluate the CoT itself using an "ideal" retriever, and the labels in the dataset are well-defined for the specific evidence passages used.

Fermi and Wikipedia evidence. We retrieve evidence against Wikipedia, in the interest of having clean and reliable evidence, and due to the fact that StrategyQA, MuSiQue and Sports Understanding are sufficiently addressed by Wikipedia knowledge. However, Fermi is a dataset which is explicitly designed to require knowledge that is difficult to attribute or measure, and require difficult inference for both knowledge and logic claims. As such, Fermi is by design not well-addressed by Wikipedia retrieval. We chose Fermi as one of our four datasets in order to have sufficient variety in the data for claims that are very difficult to support.

Reasoning chains in formats other than Chain-of-Thought. Our dataset uses CoT as a writing style for reasoning chains, due to its popularity and sentence structure that allows convenient step separation. It is of course possible to reason about complex questions in other ways, such as using sentences that combine multiple facts together, or claims that combine knowledge and logic together, and so on. Such examples can be considered a more difficult case for reasoning verification, as they will potentially require an additional solution for extracting atomic claims from the reasoning chain. CoT-format chains, such as in our dataset, do not have this requirement.

Acknowledgments

We thank Eran Ofek for his assistance in providing expert annotations, and to Idan Szpektor and Avi Caciularu for providing helpful feedback throughout the project. We are grateful to the annotators who provided labeling judgments and useful feedback throughout the annotation process.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *CoRR*, abs/2005.14165.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams

Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.

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.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics (TACL)*.

Olga Golovneva, Moya 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, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Xiaodong Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2023. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. 2023a. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks.

Alon Jacovi, Avi Caciularu, Jonathan Herzig, Roee Aharoni, Bernd Bohnet, and Mor Geva. 2023b. A comprehensive evaluation of tool-assisted generation strategies. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13856–13878, Singapore. Association for Computational Linguistics.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashwin Kalyan, Abhinav Kumar, Arjun Chandrasekaran, Ashish Sabharwal, and Peter Clark. 2021. How much coffee was consumed during EMNLP 2019? fermi problems: A new reasoning challenge for AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7318–7328, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2018. Towards automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection. *CoRR*, abs/1809.08193.

Andrew K. Lampinen, Nicholas A. Roy, Ishita Dasgupta, Stephanie C. Y. Chan, Allison C. Tam, James L. McClelland, Chen Yan, Adam Santoro, Neil C. Rabinowitz, Jane X. Wang, and Felix Hill. 2022. Tell me why! explanations support learning relational and causal structure. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 11868–11890. PMLR.

Christoph Leiter, Piyawat Lertvittayakumjorn, M. Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2022. Towards explainable evaluation metrics for natural language generation. *ArXiv*, abs/2203.11131.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y. Zhao, Yi Luan, Keith B. Hall, Ming-Wei Chang, and Yinfei Yang. 2021. Large dual encoders are generalizable retrievers.

Juri Opitz and Anette Frank. 2021. Towards a decomposable metric for explainable evaluation of text generation from AMR. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1504–1518, Online. Association for Computational Linguistics.

Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther, and Matthias Samwald. 2023. Thoughtsource: A central hub for large language model reasoning data.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *CoRR*, abs/2105.11447.

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.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021. Measuring attribution in natural language generation models. *CoRR*, abs/2112.12870.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan

Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Sean Welleck, Jiacheng Liu, Ximing Lu, Hannaneh Hajishirzi, and Yejin Choi. 2022. Naturalprover: Grounded mathematical proof generation with language models. *ArXiv*, abs/2205.12910.

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

Hugo Zaragoza, Nick Craswell, Michael Taylor, Suchi Saria, and Stephen Robertson. 2004. Microsoft cambridge at trec-13: Web and hard tracks. In *IN PROCEEDINGS OF TREC 2004*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball.

A REVEAL: Extended Details

Here we provide additional details for the collected dataset and the collection process.

A.1 Source Data Collection

Tab. 5 shows examples from each of the four source datasets, alongside example answers and their correctness. These examples are shown here for illustration, and were used as part of the pilot annotation phases, so they are not examples from the final REVEAL dataset.

To generate CoT answers, we constructed a separate prompt for each dataset, using examples from its training set with verified (correct) CoTs. StrategyQA, MuSiQue and Fermi provide gold-reference solutions (in non-CoT format) that we used to write the prompt CoT demonstrations. For Sports Understanding, we wrote the CoT demonstrations given the gold answer from the dataset. The CoT prompts used 6 demonstrations (question and CoT answer pairs) each.

CoT answers were split into sentences using NLTK's sentence tokenizer (Bird et al., 2009), and each sentence was considered one reasoning step. For purposes of retrieval they were decontextualized with the decontextualization model by Choi et al. (2021). The evidence paragraphs were retrieved from a 2021 image of Wikipedia. The evidence paragraphs in the dataset have on average 103 words, and the longest paragraph has 582 words.

A.2 Annotation Questionnaire

We split the annotation into two tasks, one focused on the logic annotation (including relevance, step type and logical correctness ratings), and the other focused on the attribution annotations (including relevance and step-evidence attribution). The annotation interfaces are shown in Fig. 8.

Task 1 is implemented and phrased as follows:

1. **Relevance:** *Is step i relevant to answering the question?* (A) Yes, it is relevant. The information in this step might be helpful to answer the question. (B) No, it's not relevant. The information in this step is not helpful to answer the question.

2. **Type:** *Does step i bring new information or describes a logical step?* (A) Attribution step. The step brings new information to help answer the question. (B) Logic step. The step only makes logical inferences based on the question or previous steps. (C) Both. The step introduces new information and makes a logical inference.

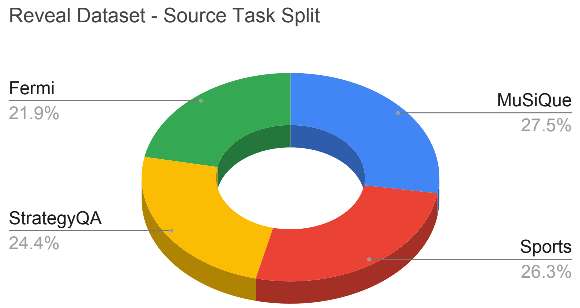
3. **Logic:** *Considering only the logical inference done in step i, is it consistent with the previous steps?* (A) Yes, correct inference. The inferences that the step makes based on the question or previous steps is correct. (B) No, incorrect inference. The inferences that the step makes based on the question or previous steps is incorrect.

4. **Justification:** *Provide a justification for your ratings.*

Task 2 is implemented and phrased as follows:

1. **Relevance:** *Should claim i be attributed?* (A) Yes, it's necessary to attribute the claim. The claim has information that needs to be verified. (B) No, it's unnecessary to attribute the claim. The claim doesn't have any information that requires verification.

2. **Attribution:** *To what extent can the information in claim i be verified by evidence j?* (A) Fully. Strictly according to the cited evidence, the claim is correct. (B) Partially. All important information in the claim is supported (no contradiction), but some information is



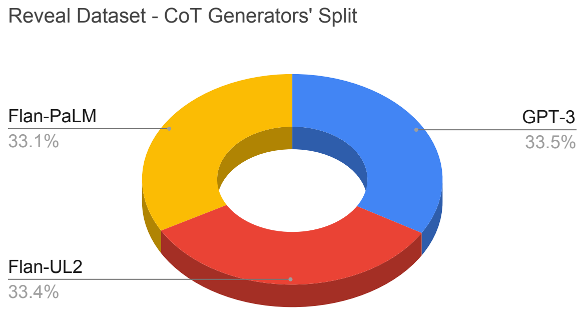


Figure 7: Distributions in REVEAL for the source tasks and CoT generators, which are approximately balanced.

Source Data	Question	Answer	Error
StrategyQA	Would a black widow have use for peaches?	A black widow is a spider. Peaches are a fruit. Thus, a black widow would have use for peaches. So the answer is yes.	Logic (step 3)
StrategyQA	Will more people go in and out of Taco Bell than a Roy Rogers each year?	Taco Bell is a fast food chain with thousands of locations. Roy Rogers is a fast food chain with fewer than 500 locations. Thus, more people will go in and out of Taco Bell than Roy Rogers each year. So the answer is yes.	Correct
MuSiQue	Who is the other child of Chian Hsiao-chang's father?	Chian Hsiao-chang's father is Chian Ching-kuo. Chian Hsiao-chang's other child is Chian Ching-wen. So the final answer is: Chian Ching-wen.	Attribution (step 2), logic (step 3)
MuSiQue	Which New York City team does not host NBA games at the site where Jimmy Carter was nominated for president?	Jimmy Carter was nominated for president at Madison Square Garden in New York City. Madison Square Garden is home to the New York Knicks, but not the Brooklyn Nets. So the final answer is: Brooklyn Nets.	Correct
Sports	Is the following sentence plausible? "Jamison Crowder changed direction in the backfield."	Jamison Crowder is a baseball player. Changing direction in the backfield is part of American football. So the answer is no.	Attribution (step 1)
Sports	Is the following sentence plausible? "Yoan Moncada beat the buzzer."	Yoan Moncada is a baseball player. Beating out the buzzer is part of basketball. So the answer is no.	Correct
Fermi	What is the volume, cc, of blood in a healthy new-born baby?	The mass of blood in a new born baby is 25 grams. The density of blood is 0.994 g*cm**-3 or grams per cc. Thus, the cc volume of blood in a new-born baby is 25g / 0.994 g*cm**-3 = 25.15 cm**3. So the final answer is: 25.15	Attribution (step 1)
Fermi	What is the total US salary?	The per capita income in the US is 63051 USD. There are around 123e+6 full time employees in the US. Thus, the total US salary is $ 63051 * 123e+6 = 7755273000000 USD. So the final answer is: 7755273000000	Correct

Table 5: Examples from each of the source datasets, alongside CoT answers, and their type of error.

unsupported. (C) Contradictory. Some information in the claim is contradicted by the cited evidence. (D) Unsupported. Some important information (or more) in the claim is not supported by the cited evidence (but no contradiction).

3. **Justification:** *Provide a justification for your attribution rating. What parts of the claim are and are not supported by the evidence?* Feel free to paste parts of the response and the source document as evidence.

A.3 Additional Statistics

Distributions across source tasks and models are shown in Fig. 7. To allow for a method of comparing multiple CoT answers to the same question, there are 161 questions which were answered by all three models in the dataset. Of them, 96 have all three CoT answers in REVEAL-Eval, while the remainder 65 questions have one or more CoT answer in REVEAL-Open.

Each step's annotation took between 300 to 600 seconds, with some answers having 5 or more steps, and with 5 annotations per step—leading to an expensive annotation process.

A.4 Analyses

Tab. 10 contains the full analysis for the properties of unsupported attribution steps.

B Experiments and Analyses

B.1 Experiment Details

In the few-shot evaluation settings, prompts were constructed by randomly selecting 8 demonstrations from class-balanced sets of 13 demonstrations. Prompts were trimmed to the biggest number of demonstrations which, combined with the query, were under the context length for the given LM.

The demonstrations were taken from the training sets of the source datasets, or from the examples used for the pilot annotation phase (which were discarded from the final REVEAL dataset).

The dataset used to train the T5-based classification baseline is a compilation of multiple datasets: MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), FEVER (Thorne et al., 2018), Sci-

Baseline	F1 (Attribution Task)	
	Fully attributable	Not fully attributable
Flan-UL2-20B	51.9	70.3
Flan-PaLM-540B	78.8	91.3
PaLM-2-L	80.0	91.9
GPT-3	74.5	88.3
FacTool	62.5	79.7
T5-xxl-true	83.0	93.8

Table 6: Baseline results for the attribution task. The models classify each step-evidence pair.

Tail (Khot et al., 2018) and VitaminC (Schuster et al., 2021).

B.2 Additional Results

Per-class F1 metrics in all settings are shown in Tables 6 to 8.

B.3 Prompt Templates

We describe here the template structures we used for our few-shot verification prompts. We note that we made an effort to test multiple templates and settings, to make sure that an adequate effort was allocated to implementing baselines that are as strong as possible.

Attribution Task

Premise: [evidence paragraph]

Hypothesis: [CoT step]

Output: {Entailment, Not Entailment}

Logic Task

Premise: [previous CoT steps]

Hypothesis: [CoT step]

Output: {Correct, Incorrect}

Step Type Task

Question: [question]

Answer: [full CoT]

Step: [CoT step]

Step type: {Attribution, Logic, Both}

Relevance Task

Question: [question]

Answer: [full CoT]

Step: [CoT step]

Is this step relevant to the answer? {Yes, No}

CoT Correctness task

Baseline	F1 (Logic Task)	
	Logically correct	Logically incorrect
Flan-UL2-20B	85.1	33.8
Flan-PaLM-540B	90.2	47.1
PaLM-2-L	90.3	64.9
GPT-3	86.7	32.2
T5-xxl-true	57.6	37.1

Table 7: Baseline results for the logical correctness task. The models classify the logical correctness of each step given the previous steps.

Baseline	F1 (Full Correctness)	
	Correct CoT	Incorrect CoT
Flan-UL2-20B *(pipeline)*	29.3	79.5
Flan-UL2-20B	29.3	54.0
Flan-PaLM-540B *(pipeline)*	31.9	84.3
Flan-PaLM-540B	40.2	38.6
PaLM-2-L *(pipeline)*	65.2	87.5
PaLM-2-L	51.4	72.4
GPT-3 *(pipeline)*	56.7	87.2
GPT-3	37.1	34.4

Table 8: Baseline results for the task of classifying the full correctness of a given CoT as an answer to a question. The pipeline variants are classifiers built on the models' answers to each sub-task (relevance, type, attribution and logic), while the non-pipeline variants simply prompt the models to classify the entire CoT.

Evidence i: [evidence paragraph]

Question: [question]

Answer: [full CoT]

The answer is: {Correct, Incorrect}

All evidences are added to the demonstration for i in the number of evidences (i.e., the number of attribution steps in the CoT).

Complexity Category	% steps in category		Description
	Attribution (N=199)	Logic (N=52)	
World knowledge or general inference	25.63%	15.38%	Some annotators did not apply world knowledge or general implicit inferences that other annotators take for granted (e.g., that a civil war takes place within a single country, or that an animal described as having prey can be called a "predator").
Acceptable nuance/hedging	1.51%	–	Disagreement between annotators on strictness of hedged approximations in the step. E.g., "There are around 7.5e+9 people in the world" where the evidence mentions "7.55 billion".
Insufficient nuance/hedging	13.07%	–	A step mentions an approximation of some quantity in the evidence but without appropriate hedging.
Averages and ranges	10.05%	–	A step picks a specific point or subrange from a more general range or set of ranges given in the evidence, and annotators disagree on whether the exemplification is representative of the evidence.
Calculation or unit issue	–	42.31%	The step consists of a calculation or unit conversion which is difficult to verify.
Specialized knowledge	13.07%	7.69%	The step requires specialized knowledge, e.g., rules of a sports game or scientific notation, which some annotations consider as world knowledge.
Formatting issue	3.02%	–	A problem in the annotation interface (e.g., truncation) made the claims impossible to verify.
Inconsistent/unclear reference or standard	12.56%	30.77%	A step makes an ambiguous reference. E.g., "Westworld was directed by Jonathan Nolan" is paired with evidence on "Westworld TV series . . . with Nolan directing the pilot episode"—Westworld could refer to the film, TV series, etc.
Inference across claims in evidence	1.51%	–	Verifying the step requires drawing an implicit conclusion, which only some annotators did—e.g., the step "Birds and mammals are different classes of animals" is paired with evidence that describes characteristics of mammals, and mentions that they "distinguish them from reptiles and birds", which implies that they are in different classes.
Invalid inference in previous step	0.50%	28.85%	Some annotators may have found it difficult to disentangle a current step's inference from any previous inferences that are invalid.
Rating category definition/criterion	23.62%	–	Disagreements stemming from the distinction between "partially attributable" and "not attributable" (despite guidance during the annotation instructions—more discussion below).
Relevance dispute	–	23.08%	Borderline-relevant steps or difficult to follow steps can lead to disagreement on the relevance label.
Temporal inconsistency	9.55%	1.92%	The step makes a claim that relies on a different time frame from that of the evidence or other steps in the answer (e.g., present-tense in "Maradona is a soccer player" where the evidence mentions "a former professional footballer"). Annotators' acceptance of temporal inconsistencies may vary (e.g., the convention of using the present tense to refer to a retired person's most salient profession).

Table 9: An overview of the different categories we surface for the disagreements observed in REVEAL-Open.

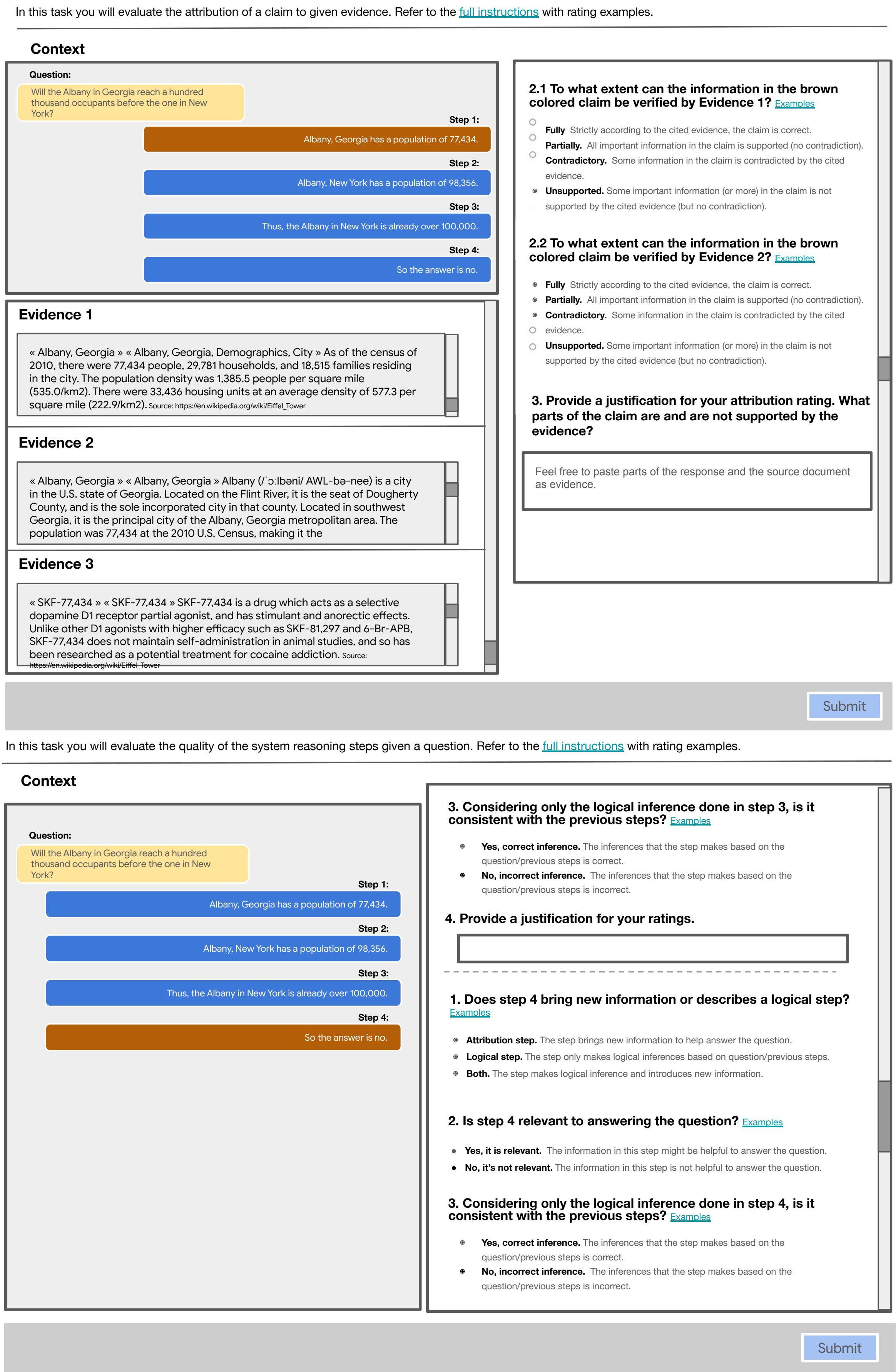


Figure 8: The annotation interface for the two tasks: Attribution (above) and logic (below).

Dataset	Step	Factually Correct?	Analysis
Fermi	There are 1000 * 100 = 10000 pennies in 1 US dollar.	No	Wrong math
Fermi	Art Beat is a free, annual event in South Bend, Indiana.	Yes	The info could be found outside of Wikipedia, the retrieved evidence was irrelevant.
Fermi	The latent heat of vaporization of steam is 540 cal/g.	Yes	The info could be found outside of wikipedia, the retrieved evidence was irrelevant.
Fermi	The total volume of the oceans is 1.3e+21 liters.	No	The total volume of the oceans is much more than 1.3e+12 liters. The retrieved evidence was relevant but did not mention the answer explicitly (metric conversion was required).
Fermi	The average American yard is 8000 square feet.	No	Could not find evidence that mentions this number, and the retrieved evidence was irrelevant.
Musique	Jayantha Ketagoda is married to his wife, Jaya.	No	Could not find evidence that mentions this claim, and the retrieved evidence was relevant to the entity but did not mention the claim.
Musique	Vibullia Alcia Agrippina's child was a proponent for the movement called the "Feminist Movement".	No	Could not find evidence that mentions this claim.
Musique	The Almost are signed to the record label called Almost Records.	No	The claim is wrong, and the retrieved evidence was not relevant. Could find evidence in wikipedia that refutes the claim.
Musique	The most Champions League wins between 1992 and 2013 were by the Spanish La Liga.	Yes	The claim is correct, but requires complex reasoning (understanding which teams are from the Spanish La Liga). The retrieved evidence was related to La Liga but not sufficient to support the claim.
Musique	Deborah Estrin is a member of the Association for Computing Machinery.	Yes	The claim is correct. There was evidence that supported the claim, but this required domain knowledge (that "ACM"="Association for Computing Machinery") so the raters marked it as unsupported.
Sports	Doing a maradona on the defender is part of basketball.	No	The claim is wrong, "a maradona" is a soccer move, not a basketball move.
Sports	Hitting the back of the rim is a common occurrence in basketball.	Yes	The claim is correct. None of the evidence supported it directly, as it is more "common knowledge".
Sports	Windmill dunk is part of basketball, not hockey.	Yes	The claim is correct, however it is a composite claim - 1) Windmill dunk is part of basketball 2) Windmill dunk is not part of hockey. The evidence supported the first claim, not the second.
Sports	Brown is a basketball player.	Yes	The claim is correct, but requires further decontextualization (from "Brown" to "Jaylen Brown") as it is vague in its current form, which resulted in irrelevant evidence.
Sports	Being out at second is part of baseball.	Yes	The claim is correct, but requires common knowledge - that "out at second" means "second base". The retrieved evidence did not support this claim explicitly.
StrategyQA	A king size bed is 76 inches wide and 80 inches long.	Yes	The step is factually correct and a google search was able to retrieve relevant evidence. The retrieved evidence was irrelevant.
StrategyQA	Pacifists do not support violence, including stoning.	Yes	The step is factually correct, and can be verified with common sense. The retrieved evidence did not support this claim.
StrategyQA	An existential crisis is a mental health issue.	Yes	The step is factually correct, and can be verified with common sense. The retrieved evidence did not support this claim explicitly.
StrategyQA	Lil Jon's top ranked Billboard song was "Get Low".	No	The claim is wrong since the song is not the only top-ranked song on Billboard from the artist. The retrieved evidence was relevant and partially supported the claim (saying that the song was indeed top-ranked in the chart).
StrategyQA	Cricketers do not kick field goals.	Yes	The claim is correct since a field goal is a term from other sports and kicking is generally prohibited in Cricket. The retrieved evidence was irrelevant since this requires more common-knowledge and inference.

Table 10: Analysis for steps where no supporting evidence was found.