Evaluating Step-by-step Reasoning Traces: A Survey

Anonymous ACL submission

Abstract

Step-by-step reasoning is widely used to enhance the reasoning ability of large language models (LLMs) in complex problems. Evaluating the quality of reasoning traces is crucial for understanding and improving LLM reasoning. However, the evaluation criteria remain highly unstandardized, leading to fragmented efforts in developing metrics and meta-evaluation benchmarks. To address this gap, this survey provides a comprehensive overview of step-by-step reasoning evaluation, proposing a taxonomy of evaluation criteria with four top-level categories (groundedness, validity, coherence, and utility). We then categorize metrics based on their implementations, survey which metrics are used for assessing each criterion, and explore whether evaluator models can transfer across different criteria. Finally, we identify key directions for future research.

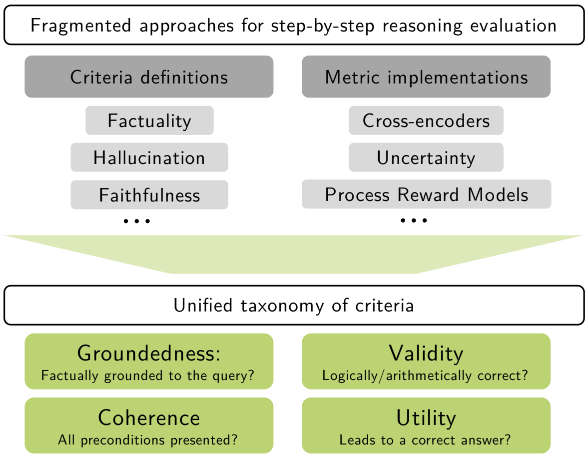

Figure 1: This survey aims to provide a comprehensive view of different terminologies on criteria and metrics designed for step-by-step reasoning evaluation.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in reasoning in complex problems, such as logic, math, and science. At the core of this versatility lies **step-by-step reasoning** (Wei et al., 2022b; Kojima et al., 2022), where the LLM generates an intermediate reasoning trace before presenting the final answer.

The step-by-step reasoning ability of LLMs is often measured in terms of *answer accuracy*, *i.e.* finding the correct answer in a problem that requires complex reasoning (OpenAI, 2024a; Groeneveld et al., 2024; DeepSeek-AI, 2025). However, answer accuracy is generally insufficient for measuring LLMs' reasoning ability, as the correct answer does not imply the correctness of the preceding reasoning trace (Lanham et al., 2023; Mirzadeh et al., 2024; Paul et al., 2024). Furthermore, the quality of the reasoning trace is crucial for improving the reasoning ability, in terms of reinforcement learning (Lu et al., 2024; Qwen-Team, 2024; DeepSeek-AI, 2025) and inference-time search (Wang et al., 2023c; Yao et al., 2023).

Due to its importance, step-by-step reasoning evaluation is a rapidly evolving field with numerous new metrics and criteria actively proposed. Establishing the precise definition of the **criterion** (*what to evaluate*) is crucial for correctly implementing the **metric** (*how to evaluate*). However, the terminologies in the field are highly unstandardized, which has led to fragmented approaches in implementing metrics and meta-evaluation benchmarks. This current state motivates a systematic review, which will serve as a foundation for general criteria and metrics that can span diverse reasoning tasks.

In this survey, we reorganize existing step-by-step reasoning evaluation criteria defined within diverse metrics and meta-evaluation benchmarks into four distinct categories: factual **groundedness** in the given information, logical **validity** of steps, semantic **coherence**, and if the step contributes to the correct answer (**utility**). Based on the proposed taxonomy, we review and compare widely used terms for criteria and metrics. Finally, we analyze the case of *transferability*, whether a single evaluator trained/optimized for one criterion can evaluate another, based on reported scores on three recent meta-evaluation benchmarks (Jacovi et al., 2024; Song et al., 2025; Zheng et al., 2024). Finally, we conclude the survey with open questions in the field of evaluating step-by-step reasoning.

The key contributions of this survey are:

- Defining the taxonomy of step-by-step evaluation **criteria**, and comparing it with existing terminologies (§3-§4).
- Surveying existing **metrics** for step-by-step reasoning evaluation based on their implementations, across diverse reasoning tasks and criteria (§5).
- Analyzing **transferability** between criteria based on reported empirical results (§6).

2 Background

2.1 Step-by-step reasoning evaluation

Step-by-step reasoning is where LLMs generate a series of intermediate natural language steps that lead to the final answer (Wei et al., 2022b). Each step-by-step reasoning consists of three parts, a **query**, a **reasoning trace**, and the **answer** (Figure 2). Query refers to the entire input, which includes the question and retrieved evidence in fact-intensive reasoning tasks (Lewis et al., 2020). Upon seeing a query, the LLM autoregressively generates its solution as a long **reasoning trace**. Finally, a trace should output an **answer**, either explicitly formatted (*e.g.* `\boxed{15}`) or implicitly stated (*e.g. Therefore, John ate 15 apples*).

Various evaluation metrics require the reasoning trace to be segmented into **steps**. The step boundary can be determined using simple rules, *e.g.* sentences or double newlines (`\n\n`). However, the format of a reasoning trace is highly dependent on the format of the instruction tuning data, which might lead to inconsistent granularity of steps. As a solution, alternative segmentation strategies were proposed, including Semantic Role Labeling-based chunking Prasad et al. (2023) or prompting LLMs Zheng et al. (2024).

Finally, metrics assess the quality of the step and assign a **score**. The details about different metrics are further described in Section 5. These scores can be used to improve answer accuracy in Best-of-N decoding (Cui et al., 2024; Zhang et al., 2025),

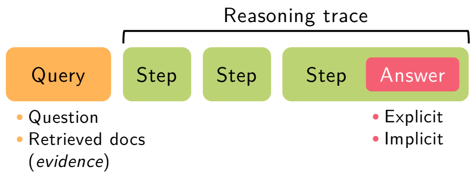

Figure 2: Illustration of three elements of step-by-step reasoning: query, reasoning trace (steps), and the answer.

train LLMs via reinforcement learning (Wang et al., 2024b; Zhang et al., 2025), or guide inference-time tree search (Yao et al., 2023; Yang et al., 2022).

2.2 Reasoning tasks

The concept of step-by-step reasoning was initially derived from **factual/commonsense reasoning**. These tasks include questions that can only be answered by combining different information from the query and performing multi-hop inference (Mavi et al., 2024). **Factual reasoning** focuses on combining facts to find the correct answer (Yang et al., 2018; Talmor and Berant, 2018; Kwiatkowski et al., 2019), while **commonsense reasoning** also requires commonsense knowledge to complete the inference (Clark et al., 2018; Talmor et al., 2019; Geva et al., 2021; Trivedi et al., 2022).

Another important venue is **symbolic reasoning**, where the reasoning process can be expressed using *symbols* (*e.g.* equations, logic, code) (Sprague et al., 2024). This encompasses **mathematical reasoning**, including arithmetics, calculus, and number theory (Cobbe et al., 2021; Hendrycks et al., 2021; He et al., 2024a; Gao et al., 2024b); **logical reasoning**, which involves performing complex sequence of deductive inference (Tafjord et al., 2021; Han et al., 2024a; Saparov and He, 2023); and **algorithmic reasoning**, which requires manipulating strings or data structures (BIG-Bench-Team, 2023; Suzgun et al., 2022; Valmeekam et al., 2023).[1]

Further details on reasoning tasks and benchmarks are presented in Appendix A.

3 Taxonomy

This section aims to provide a clear taxonomy of criteria for evaluating step-by-step reasoning. Existing criteria can be seen as falling into one of the four categories, namely **Groundedness**, **Validity**,

[1]While symbolic reasoning may strictly refer to *algorithmic reasoning* (Wei et al., 2022b) depending on context, we adopt the broader sense that includes math and logical reasoning.(Sprague et al., 2024).

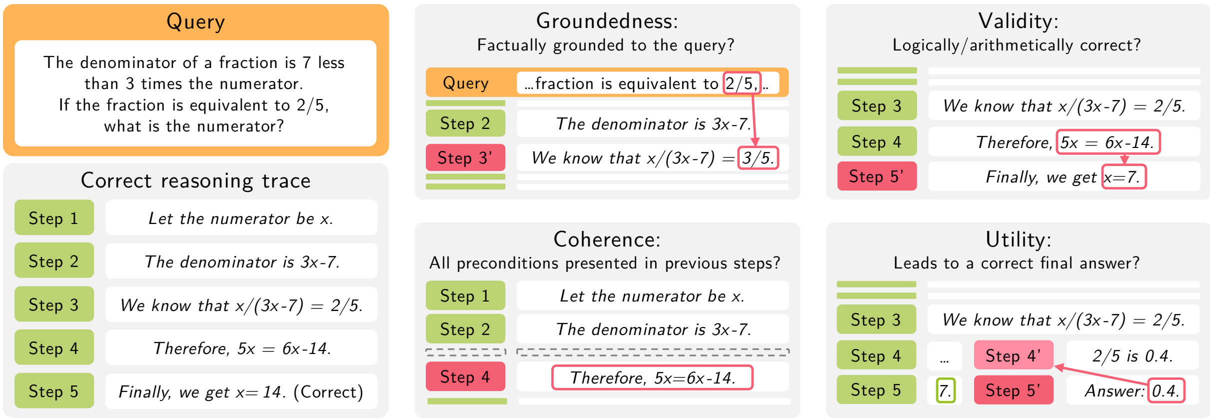

Figure 3: Illustration of the proposed categories of step-by-step reasoning evaluation criteria, *i.e.* groundedness, validity, coherence, and utility. The left shows an example of a query and a reasoning trace. The other four blocks demonstrate examples that fail to suffice the respective metric. Red filled rectangles indicate the error's location, and the outlined boxes and arrows show the cause of the error.

Coherence, and **Utility**. These definitions are *independent* (aim at different objectives – Section 4.1), but *not mutually exclusive* (a step can fail to suffice multiple criteria at once).

3.1 Groundedness

Groundedness evaluates if the *step is factually true* according to the query (Lewis et al., 2020; Gao et al., 2024d). A step can be ungrounded to any part of the query, *e.g.* the question (Figure 3-Groundedness) or evidence (*e.g.* falsely stating that *Buddy Rich was born in Chicago*, where the retrieved document states that he was born in New York).

3.2 Validity

Validity evaluates if a reasoning step contains no errors.

The validity of a reasoning step can be defined in terms of *entailment* (Bowman et al., 2015), which is widely accepted in factual/commonsense reasoning. Under this definition, a step is considered valid if it can be directly entailed from previous steps (Tafjord et al., 2021; Dalvi et al., 2021; Saparov and He, 2023) or at least does not contradict them (Golovneva et al., 2023a; Prasad et al., 2023; Zhu et al., 2024b).

The notion of validity often used in symbolic tasks is *correctness*, *e.g.* performing accurate calculations in math reasoning (Lightman et al., 2024; Jacovi et al., 2024; Zheng et al., 2024) or inferring the correct logical conclusion based on the provided premises (Wu et al., 2024b; Jacovi et al., 2024; Song et al., 2025).

3.3 Coherence

Coherence measures if a reasoning step's *preconditions are satisfied* by the previous steps (Wang et al., 2023a). For instance, if a trace includes the reasoning step *"Next, we add 42 to 16."* but the origin of the value 42 was never explained in the previous steps, this step is considered incoherent. An intuitive way to obtain an incoherent trace is randomly shuffling a coherent trace (Wang et al., 2023a; Nguyen et al., 2024), as the premise of some steps will not appear anywhere in the previous steps even though it can be eventually deduced (*valid*).

Note that coherence judgment is inherently subjective and pragmatic compared to other criteria. For instance, seemingly trivial steps like *"A part of something is present in that something"* in WorldTree V2 (Xie et al., 2020) is annotated as necessary in Dalvi et al. (2021) but not necessary in Ott et al. (2023).

3.4 Utility

Utility measures whether a reasoning step contributes to getting the correct final answer (*answer correctness*).

One interpretation of utility is *progress*, or whether the step is correctly following the ground truth solution. For instance, in Game of 24 (making the number 24 using 4 natural numbers and basic arithmetic operations) (Yao et al., 2023), a solution can be defined as a sequence of operations (*e.g.* 5+7=12→12-6=6→6*4=24.). In this task, the utility of a step (making $5 + 7 = 12$ from 5 and 7) can be directly assessed by checking if it is a part of a correct solution.

Utility can also be interpreted as *value function* (estimated reward), which is proportional to the probability of reaching the correct answer starting from the step (Hao et al., 2023; Wang et al., 2024b; Xie et al., 2024; Chen et al., 2023). This black-box interpretation of utility offers high scalability as it only requires the gold answer, without any human annotation or ground-truth solutions (Wang et al., 2024b; Lai et al., 2024).

4 Comparative analysis

4.1 Comparison between proposed categories

Groundedness↔Validity. Groundedness focuses on the explicit information in the query while validity focuses on the inference. For instance, Given an incorrect step *Albert Einstein died in 1965* (he died in 1955), this step is not grounded if the query explicitly mentions that *Einstein died in 1955*. Apart from that, if the previous steps provide the premises for reaching 1955, *i.e. Einstein was born in 1879, and he died at the age of 76*, the step is invalid.

Validity↔Coherence. Existing works often treat coherence as a subtype of validity (Golovneva et al., 2023a; Zhu et al., 2024b; Kim et al., 2024b; Jacovi et al., 2024), as both criteria judge a step based on its previous steps. However, validity and coherence are different by definition, as validity focuses on the logical correctness of a step while coherence focuses on the pragmatic aspect of informativeness. For instance (Figure 3-Coherence), omitting a step (Step 3) from the correct trace will make the subsequent step (Step 4) incoherent, but Step 4 is still valid since it can be eventually deduced from the query and previous steps.

Validity↔Utility. Previous studies have continuously pointed out that validity does not necessarily lead to utility and vice versa (Lyu et al., 2023; Nguyen et al., 2024). One case is *shortcut reasoning* (Schnitzler et al., 2024; Lee and Hwang, 2025), where LLM generates invalid Chain-of-thoughts but guesses the correct answer directly from the query. ProcessBench (Zheng et al., 2024) reports that invalid traces with correct answers can be easily found in challenging problems, reaching 51.8% in the olympiad-level Omni-MATH (Gao et al., 2024b).

4.2 Comparison to existing terminologies

Factuality is often defined as *"model's capability of generating contents of factual information, grounded in reliable sources"* (Wang et al., 2023b, 2024c), which originates from other text generation tasks such as abstractive summarization. However, this definition fails to include groundedness to the *question*, *e.g.* using the exact numbers provided in the math problem (Zhu et al., 2024b).

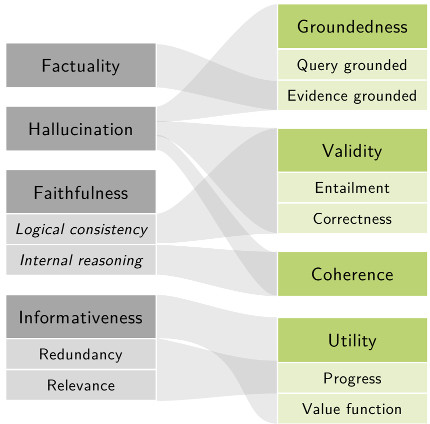

Figure 4: A Sankey diagram displaying the relationship between commonly used terminologies (left) to the proposed taxonomy (right).

Hallucination is most commonly defined as *"models either generating (1) nonsensical or (2) unfaithful to the source content"* (Ji et al., 2023; Banerjee et al., 2024; Huang et al., 2024), which corresponds to (1) validity/coherence and (2) groundedness. However, some works restrict the meaning of hallucination to groundedness errors, *i.e. "models generating description tokens that are not supported by the source inputs"* (Xiao and Wang, 2021; Akbar et al., 2024).

Faithfulness is also used in different contexts. The most common definition for faithfulness is *"logical consistency between the generated text and the query/previous steps"* (Maynez et al., 2020; Creswell and Shanahan, 2022; Huang et al., 2024), which includes both groundedness (query) and validity (previous step). Instead, faithfulness can be used as *"accurately representing the model's internal reasoning process"* (Lyu et al., 2023; Lanham et al., 2023). Under this definition, the final step containing the answer is unfaithful if it is not supported by the previous steps, which falls under the definition of coherence.

Informativeness is defined as "providing new information that is helpful towards deriving the generated answer" (Golovneva et al., 2023b; Prasad et al., 2023). Lack of informativeness is often described as **redundancy** *"removing the step does not affect the reasoning process"* (Chiang and Lee, 2024; Song et al., 2025; Zhou et al., 2024) or **ir-**

relevance *"unrelated to the query's topic or task"* (Wang et al., 2023a; Zhou et al., 2024; Jacovi et al., 2024). Informativeness is highly related to utility, as it aims to evaluate the contribution of a step to reaching the final answer.

5 Metric implementations

Metric impl.	G	V	C	U
Rule-based	▲	▲	▲	▲
Uncertainty	●			▲
$\mathcal{V}$-information		▲		●
Cross-encoder	●	●	▲	▲
PRM	▲	●	▲	●
Critic models	●	●	●	●
Generative verifiers		●		
LLM-as-value-function				●

Table 1: Mapping between each metric implementation type to the category commonly used. For each combination of metric and implementation, ● denotes that there are at least 3 published works, and ▲ denotes that there are 1 or 2. The full table can be found in Appendix C.

Numerous metrics have been proposed to evaluate and quantify the quality of a reasoning trace beyond the answer correctness. This section provides an overview of these methods, from rule-based metrics to neural models.

5.1 Rule-based matching

For tasks where the ground truth solution can be expressed as a *graph of entities*, one can view a step as a directed edge between two entities. Typical examples include knowledge graphs for factual reasoning Nguyen et al. (2024) or computation graphs in arithmetic problems (Li et al., 2023). In this setting, groundedness corresponds to having the necessary entities given in the query, validity to predicting the relation between entities, coherence to the correct ordering of steps, and utility to the existence of the step in the gold reasoning chain (Nguyen et al., 2024; Saparov and He, 2023). However, this approach may not generalize well for tasks that do not have a straightforward graph representation, *e.g.* commonsense reasoning or complex math reasoning beyond arithmetic word problems.

5.2 Intrinsic properties

Uncertainty. Uncertainty of the model can be used as an intrinsic proxy about the generated content's quality (Xiao and Wang, 2021; Zhang et al., 2023b). Qiu et al. (2024) and Wu et al. (2024a) use *token probability entropy* (Figure 5(a)), defined as $\Sigma_{t\in V} p(t)\log(p(t))$ where p is the probability distribution of all tokens in vocabulary V. Farquhar et al. (2024) and Kossen et al. (2024) extend the approach by clustering semantically similar answers and calculating the entropy with respect to the clusters. Another variant of uncertainty uses **confidence**, *i.e.* $\max_{t\in V} p(t)$ (Wu et al., 2024a; Wang et al., 2024d). In this setting, higher confidence implies that the step is more grounded/correct.

$\mathcal{V}$-information. (Chen et al., 2023; Prasad et al., 2023) use **Conditional $\mathcal{V}$-information (CVI)** (Hewitt et al., 2021) to evaluate reasoning traces. CVI can be informally defined as the amount of information the evaluation target text t adds to the model. Formally, given a model g trained to predict the answer *with* t (calculates $g(a \mid q, t)$) and g' trained to predict the answer a *without* t (calculates $g'(a \mid q)$), the CVI is calculated by

$$CVI(t \to a \mid q) = -\log g'(a \mid q) + \log g(a \mid q, t)$$

which is maximized when predicting the answer without the target is hard (smaller $g'(a \mid q)$) but it becomes easier with the target (larger $g(a \mid q, t)$) (Figure 5(b)). While this definition directly corresponds to utility (Chen et al., 2023), Prasad et al. (2023) leverages CVI to evaluate validity in an ensemble with cross-encoders (introduced below).

5.3 Neural evaluator models.

Cross-encoders. Cross-encoders are neural models that simultaneously encode two sentences using a single network (Figure 5(c)). They have been widely applied to solve tasks such as natural language inference (Bowman et al., 2015) and fact verification (Thorne et al., 2018), where one has to determine if the *hypothesis* can be inferred from the given *premise*. Cross-encoders trained on these off-the-shelf tasks are used to evaluate a reasoning step based on the query (groundedness) or previous steps (validity) (Wu et al., 2024a; Zha et al., 2023; Prasad et al., 2023). Instead of using an off-the-shelf model, Zhu et al. (2024b) perturbs correct traces with LLMs and uses the synthetic data to train the cross-encoder.

Process reward models. While process reward model (PRM) is defined as *"a model that provides feedback/evaluation for each step"* in the broadest sense, in practice, it commonly refers to an LLM with a lightweight head attached to the final layer and trained to predict a numeric score in a supervised manner (Lightman et al., 2024; Wang et al.,

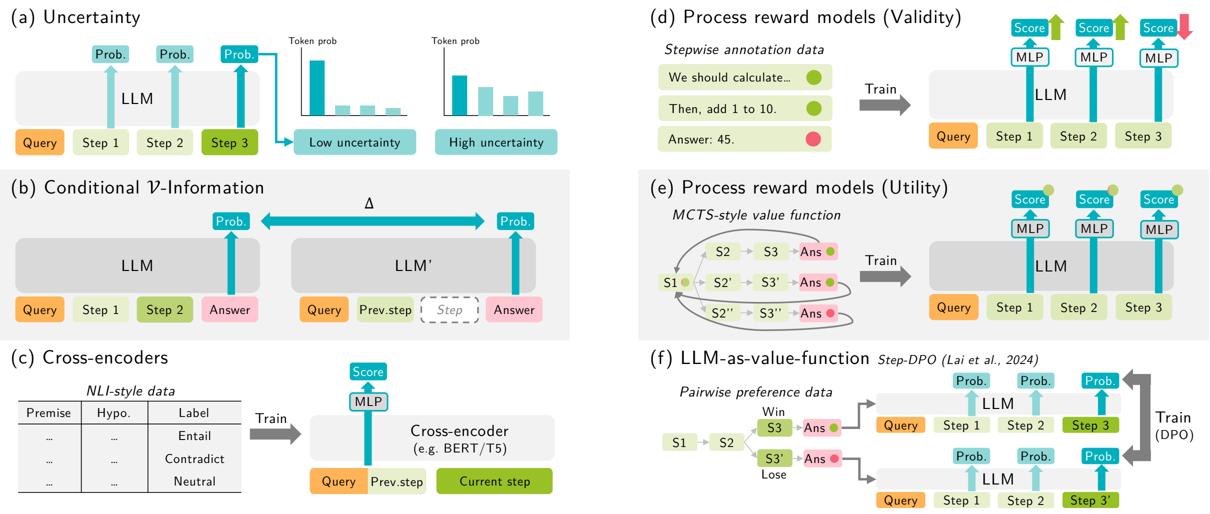

Figure 5: Illustration of six representative metric implementations. (a) and (b) use the token probabilities of the LLM generating the trace, and (c)-(e) train a separate evaluator model. (f) trains the LLM so that the token probabilities can be interpreted as scores.

2024b; Setlur et al., 2024). The training data can be categorized as (1) *validity data* including correctness annotations for each step (Hendrycks et al., 2021) (Figure 5(d)), or (2) *utility data* (Wang et al., 2024b) providing the value function obtained from Monte Carlo Tree Search (MCTS) and its variants (Figure 5(e)). We discuss the difference and transferability between these PRMs in Section 6.3.

Critic models (LLM-as-a-judge). LLM-as-a-judge (Zheng et al., 2023; Kim et al., 2024a) is a widely accepted paradigm for evaluate long texts. In reasoning trace evaluation, the term *critic models* often refers to the same concept (Zheng et al., 2024; Lin et al., 2024). Jacovi et al. (2024); Wu et al. (2024d); Niu et al. (2024); Yao et al. (2023) showed that prompting instruction-tuned LLMs can effectively evaluate groundedness, validity, coherence, and utility in diverse reasoning tasks with Chain-of-thoughts prompting (Wei et al., 2022b). The specific format of evaluation can vary from (1) evaluating if the entire trace is correct or not, (2) finding the location of the first erroneous step given the entire trace, or (3) judging a single step's correctness based on the query and previous steps.

Generative Verifiers. This paradigm lies in the middle ground of PRMs and critic models, by first generating the evaluation rationale and then using a small head to predict the numerical scores conditioned on the self-generated rationales (Ankner et al., 2024; Zhang et al., 2024b).

LLM-as-value-function. LLMs can be directly trained to align sequence probabilities (relative to the initial model's probability) to the value function as shown in Direct Preference Optimization (DPO; Rafailov et al. (2023)) (Figure 5(f)). Consequently, LLMs trained to distinguish traces with correct answers from incorrect ones by DPO can directly serve as a utility evaluator (Mahan et al., 2024; Lai et al., 2024; Xie et al., 2024; Pang et al., 2024; Cui et al., 2025), where the relative sequence probability is the utility score. Unlike PRMs that are not fine-tuned for generation, these models retain (and improve) the ability to generate. However, these models require an additional forward pass to obtain the initial model probability, doubling the computation cost during the evaluation phase.

6 Analysis on reported meta-evaluation

Based on the taxonomy provided in Section 3, we observe that a *single* evaluator model, with identical implementation design, model, and training data/prompt, is often used to evaluate different metrics. For instance, a single cross-encoder model is used to evaluate the groundedness and validity in Golovneva et al. (2023b); Zhu et al. (2024b).

However, such *transferability*, *i.e.* an evaluator tuned for one metric being able to generalize to another, is not trivial because the criteria definitions are independent. Transferability is important in terms of designing metrics and meta-evaluation benchmarks, as (*metric*) using the same model for evaluating non-transferable criteria will lead to sub-optimal performance, and (*meta-evaluation benchmark*) annotating non-transferable errors as same

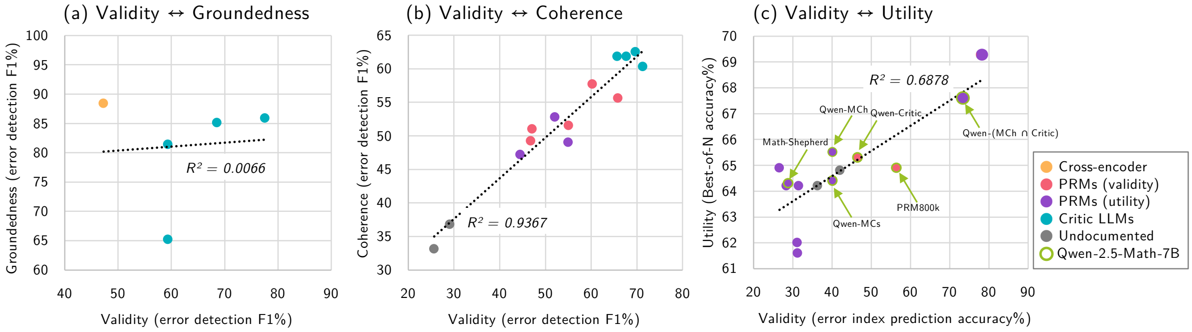

Figure 6: Meta-evaluation scores of the same evaluator model in two different criteria. (a) Results from REVEAL Jacovi et al. (2024) show that validity and groundedness are not transferrable, and cross-encoders fall behind critic models in evaluating validity. (b) PRMBench Song et al. (2025) shows that validity and coherence evaluation are highly transferable. (c) Zhang et al. (2025) shows that utility-based PRMs often fail to evaluate validity, but the two criteria can synergize when jointly considered.

categories might disrupt the meta-evaluation results. Note that high correlation does not imply that the criteria are *duplicates*, as their definition significantly differ (Section 3).

We investigate if there is evidence of transferability between criteria proposed in Section 3 by analyzing reported empirical results in three meta-evaluation settings, namely REVEAL (Jacovi et al., 2024), PRMBench (Song et al., 2025), and Process-Bench + BoN decoding (Zhang et al., 2025).

6.1 Validity-Groundedness

REVEAL (Jacovi et al., 2024) is a meta-evaluation benchmark based on commonsense reasoning. It evaluates a cross-encoder model (Honovich et al., 2022) and various critic models (LLM-as-a-judge) (Brown et al., 2020; Wei et al., 2022a; Anil et al., 2023) upon reasoning traces sampled from four commonsense reasoning benchmarks. The results (Figure 6(a)) show that the correlation between the two scores is weak, indicating that using a single model for both methods can result in suboptimal evaluation performance.

Notably, the cross-encoder model (Figure 6(a) ●) achieves significant accuracy in groundedness but falls over 10p behind critic models in evaluating validity. This result indicates that it might not be feasible to employ off-the-shelf cross-encoders trained on NLI tasks for validity judgments, as opposed to existing works (Golovneva et al., 2023a; Prasad et al., 2023).

6.2 Validity-Coherence

PRMBench (Song et al., 2025) defines nine fine-grained error classes in the PRM800k dataset (Lightman et al., 2024) and annotates $\tilde{1}50$ samples per class for meta-evaluation. Among the nine classes, we display the correlation between Step Consistency (SC; *Are the two steps contradictory?*) representing the validity error and Prerequisite Sensitivity (PS; *Are any critical premises, assumptions, or necessary conditions absent?*) representing coherence. The results (Figure 6(b)) show that the correlation is high in diverse PRMs and critic models, indicating that the abilities to evaluate validity and coherence are very likely transferable.

6.3 Validity-Utility

Recent works on process reward models do not explicitly disambiguate between validity-based and utility-based PRMs. Consequently, training the model with one data (*e.g.* validity) and evaluating with another (utility) has settled as a common experimental practice (Lightman et al., 2024; Ma et al., 2023; Zheng et al., 2024; Song et al., 2025).

In this setting, we analyze results on Process-Bench (Zheng et al., 2024) and Best-of-N decoding results, reported by Zhang et al. (2025). Process-Bench is a meta-evaluation benchmark constructed from human annotations on validity. In contrast, Best-of-N decoding tests the ability of an evaluator to select the reasoning trace with the highest utility (chance of answer correctness) out of N samples.

In Figure 6(c), the correlation between two criteria is weaker than validity-coherence ($R^2 = 0.69$). Furthermore, Zhang et al. (2025)'s analyses show that if only comparing validity and utility PRMs trained on the same base model (Qwen-2.5-MAth-7B (Yang et al., 2024), models trained on utility[2]

[2]Figure 6(c) Math-Shepherd, Qwen-MCh, Qwen-MCs

achieve significantly lower performance in validity evaluation than validity PRMs[3]. They show that filtering the training samples with high validity *and* utility scores leads to powerful PRM[4]. These results indicate that validity and utility are complementary, and considering both yields more robust evaluation results than using single criterion.

7 Future directions

Despite rapid progress on step-by-step reasoning evaluation, crucial questions remain to be solved.

Resources for evaluating reasoning in challenging real-world reasoning tasks. Datasets for training and evaluating neural reasoning trace evaluators are generally restrained to tasks that are either overly simple (*e.g.* popular MHQA datasets) or restricted in domains (*e.g.* olympiad-level math reasoning). However, there are many real-world reasoning tasks such as complex science questions (Rein et al., 2024), repository-level coding (Zhang et al., 2023a), medicine (Savage et al., 2024), law (Holzenberger and Van Durme, 2021; Kim et al., 2024c), and finance (Li et al., 2024b). The reasoning required for these tasks is complex, requiring both groundedness to retrieved documents and expert-level mathematic/logical skills. Developing step-by-step reasoning evaluators and meta-evaluation benchmarks for such expert-level tasks will significantly enhance the generalizability and real-world applicability of LLM reasoning.

Evaluation of long, complex reasoning traces. Due to the recent attention to OpenAI `o1` (OpenAI, 2024b), numerous models have been trained to generate a long reasoning trace that includes hesitation, backtracking, and lookahead assumptions (OpenAI, 2024b; Zhao et al., 2024; DeepSeek-AI, 2025; Muennighoff et al., 2025). However, existing step-by-step evaluation reasoning metrics are not designed to accommodate these complex traces. For instance, incorrect steps followed by correct self-correction (*e.g. Wait, this reasoning is not correct.*) will get low validity and utility scores because the step will lead to a contradiction and is semantically irrelevant to the final answer. While the necessity of trace evaluation in obtaining stronger long-trace models is under debate (DeepSeek-AI, 2025), the effort to develop evaluation resources for such trace will lead to a better understanding of long-trace models' behaviors and further improvement in reasoning performance.

Symbol-grounded evaluation of reasoning traces. Reasoning tasks often have a symbolic ground truth solution. For instance, deductive reasoning tasks can be represented with formal logic, and arithmetic problems can be expressed as a series of equations or symbolic theorems. These solutions provide precise, formal ways to define metrics, including validity and utility (progress). However, not much work has been done to exploit the parallel between reasoning traces and the underlying symbolic solution. While several rule-based approaches parse reasoning traces for evaluation in relatively easier reasoning tasks (Saparov and He, 2023; Nguyen et al., 2024; Li et al., 2023), no attempts have been made to extend this paradigm to evaluate reasoning traces for first-order logic reasoning (Han et al., 2024a,b) and math problems formalized using theorem provers, *e.g.* Lean (Yang et al., 2023; Gao et al., 2024c).

Objective metrics for coherence evaluation. LLMs often omit trivial inference steps in their reasoning (Saparov and He, 2023), but there is no consensus about to what extent can the step be omitted (Section 3.3). This widespread ambiguity led to a deprivation of objective coherence evaluation metrics. A large-scale annotation of omittable and non-omittable steps will facilitate the development of precise coherence evaluators and comprehensive meta-evaluation based on human perception of coherence.

8 Conclusion

This survey aims to organize the scattered terminologies and methods for step-by-step reasoning evaluation, which is crucial for understanding and improving LLM's reasoning capabilities. This survey provides a unified taxonomy for evaluation criteria, a comprehensive review on existing metrics and their implementation, and tackle transferability between different metrics.

Still, there are diverse challenges left in the field of evaluating step-by-step reasoning. As the reasoning trace becomes longer and more complex to solve challenging problems, existing methods might fail to capture the complex structure of the solution. As the step-by-step reasoning performance and trustworthiness of LLMs improve, proper and careful evaluation will surely remain crucially important.

[3]Figure 6(c) PRM800K, Qwen-Critic

[4]Figure 6(c) Qwen-MCh∩Critic

9 Limitation

This survey aims to provide a comprehensive view of step-by-step evaluation reasoning by focusing on criteria definition and metric implementations. In return, this work does not fully address the role of *human judgments* in the task, including the human annotation process (Lightman et al., 2024; Zheng et al., 2024; Song et al., 2025), human correlation (Zha et al., 2023; Golovneva et al., 2023a; Prasad et al., 2023), and inter-annotator agreement (Jacovi et al., 2024). Furthermore, while this work analyzes reported empirical results in Section 6, it does not perform additional experiments to compare more diverse metrics in a fair and comprehensive setting.

References

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica M Salinas, Victor Alvarez, and Erwin Cornejo. 2024. HalluMeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037, Miami, Florida, USA. Association for Computational Linguistics.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

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. *Preprint*, arXiv:2305.10403.

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. 2024. Critique-out-loud reward models. *Preprint*, arXiv:2408.11791.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. Llms will always hallucinate, and we need to live with this. *Preprint*, arXiv:2409.05746.

BIG-Bench-Team. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

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. *Preprint*, arXiv:2005.14165.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Lilian Weng, and Aleksander Madry. 2024. Mle-bench: Evaluating machine learning agents on machine learning engineering. *CoRR*, abs/2410.07095.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Cheng-Han Chiang and Hung-yi Lee. 2024. Over-reasoning and redundant calculation of large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 161–169, St. Julian's, Malta. 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. *Preprint*, arXiv:1803.05457.

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. *Preprint*, arXiv:2110.14168.

Antonia Creswell and Murray Shanahan. 2022. Faithful reasoning using large language models. *Preprint*, arXiv:2208.14271.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. 2025. Process reinforcement through implicit rewards. *Preprint*, arXiv:2502.01456.

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.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2024. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *Preprint*, arXiv:2312.14890.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, Junjie Hu, Tianyu Liu, and Baobao Chang. 2024a. Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. *Preprint*, arXiv:2406.14024.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. 2024b. Omni-math: A universal olympiad level mathematic benchmark for large language models. *Preprint*, arXiv:2410.07985.

Guoxiong Gao, Yutong Wang, Jiedong Jiang, Qi Gao, Zihan Qin, Tianyi Xu, and Bin Dong. 2024c. Herald: A natural language annotated lean 4 dataset. *Preprint*, arXiv:2410.10878.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024d. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

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

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *Preprint*, arXiv:2411.04872.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023a. 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.

Olga Golovneva, Sean O'Brien, Ramakanth Pasunuru, Tianlu Wang, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023b. Pathfinder: Guided search over multi-step reasoning paths. *Preprint*, arXiv:2312.05180.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. 2024a. FOLIO: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.

Simeng Han, Aaron Yu, Rui Shen, Zhenting Qi, Martin Riddell, Wenfei Zhou, Yujie Qiao, Yilun Zhao, Semih Yavuz, Ye Liu, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Dragomir Radev, Rex Ying, and Arman Cohan. 2024b. P-FOLIO: Evaluating and improving logical reasoning with abundant human-written reasoning chains. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16553–16565, Miami, Florida, USA. Association for Computational Linguistics.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024a. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.

Mingqian He, Yongliang Shen, Wenqi Zhang, Zeqi Tan, and Weiming Lu. 2024b. Advancing process verification for large language models via tree-based preference learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2086–2099, Miami, Florida, USA. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Holzenberger and Benjamin Van Durme. 2021. Factoring statutory reasoning as language understanding challenges. 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 2742–2758, Online. Association for Computational Linguistics.

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 Second*

DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4615–4634, Bangkok, Thailand. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? *Preprint*, arXiv:2310.06770.

Liwei Kang, Zirui Zhao, David Hsu, and Wee Sun Lee. 2024. On the empirical complexity of reasoning and planning in llms. *Preprint*, arXiv:2404.11041.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *Preprint*, arXiv:2406.05761.

Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. 2024c. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA. Association for Computational Linguistics.

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.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *Preprint*, arXiv:2406.15927.

Eldar Kurtic, Amir Moeini, and Dan Alistarh. 2024. Mathador-LM: A dynamic benchmark for mathematical reasoning on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17020–17027, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Xin Lai, Zhuotao Tian, Yukang Chen, Senqiao Yang, Xiangru Peng, and Jiaya Jia. 2024. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *Preprint*, arXiv:2406.18629.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, 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. *Preprint*, arXiv:2307.13702.

Jinu Lee and Wonseok Hwang. 2025. Symba: Symbolic backward chaining for structured natural language reasoning. *Preprint*, arXiv:2402.12806.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Ruosen Li, Zimu Wang, Son Tran, Lei Xia, and Xinya Du. 2024a. Meqa: A benchmark for multi-hop event-centric question answering with explanations. In *Advances in Neural Information Processing Systems*, volume 37, pages 126835–126862. Curran Associates, Inc.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024b. AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 773–783, Torino, Italia. ELRA and ICCL.

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.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning. *Preprint*, arXiv:2402.14809.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023. Deductive verification of chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 36, pages 36407–36433. Curran Associates, Inc.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Step-controlled dpo: Leveraging stepwise error for enhanced mathematical reasoning. *Preprint*, arXiv:2407.00782.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.

Qianli Ma, Haotian Zhou, Tingkai Liu, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2023. Let's reward step by step: Step-level reward model as the navigators for reasoning. *Preprint*, arXiv:2310.10080.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *Preprint*, arXiv:2410.12832.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop question answering. *Preprint*, arXiv:2204.09140.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *Preprint*, arXiv:2005.00661.

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. *Preprint*, arXiv:1809.02789.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *Preprint*, arXiv:2410.05229.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

Thi Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2862–2883, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

Skywork o1 Team. 2024. Skywork-o1 open series. `https://huggingface.co/Skywork`.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

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. *Scientific Data*, 10(1).

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *Preprint*, arXiv:2404.19733.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15012–15032, Miami, Florida, USA. Association for Computational Linguistics.

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.

Zexuan Qiu, Zijing Ou, Bin Wu, Jingjing Li, Aiwei Liu, and Irwin King. 2024. Entropy-based decoding for retrieval-augmented large language models. *Preprint*, arXiv:2406.17519.

Qwen-Team. 2024. QwQ: Reflect Deeply on the Boundaries of the Unknown — qwenlm.github.io. https://qwenlm.github.io/blog/qwq-32b-preview/. [Accessed 13-02-2025].

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations*.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Julian Schnitzler, Xanh Ho, Jiahao Huang, Florian Boudin, Saku Sugawara, and Akiko Aizawa. 2024. Morehopqa: More than multi-hop reasoning. *Preprint*, arXiv:2406.13397.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024. Rewarding progress: Scaling automated process verifiers for llm reasoning. *Preprint*, arXiv:2410.08146.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *Preprint*, arXiv:2501.03124.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *Preprint*, arXiv:2409.12183.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

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.

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.

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.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic. 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, 10:539–554.

Nemika Tyagi, Mihir Parmar, Mohith Kulkarni, Aswin Rrv, Nisarg Patel, Mutsumi Nakamura, Arindam Mitra, and Chitta Baral. 2024. Step-by-step reasoning to solve grid puzzles: Where do LLMs falter? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19898–19915, Miami, Florida, USA. Association for Computational Linguistics.

Gladys Tyen, Hassan Mansoor, Victor Carbune, Peter Chen, and Tony Mak. 2024. LLMs cannot find reasoning errors, but can correct them given the error location. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13894–13908, Bangkok, Thailand. Association for Computational Linguistics.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2023. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. In *Advances in Neural Information Processing Systems*, volume 36, pages 38975–38987. Curran Associates, Inc.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023b. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *Preprint*, arXiv:2310.07521.

Jianing Wang, Qiushi Sun, Xiang Li, and Ming Gao. 2024a. Boosting language models reasoning with chain-of-knowledge prompting. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4958–4981, Bangkok, Thailand. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. 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.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024c. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA. Association for Computational Linguistics.

Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024d. Chain-of-probe: Examing the necessity and accuracy of cot step-by-step. *Preprint*, arXiv:2406.16144.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. 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.

Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. 2024a. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9390–9406, Miami, Florida, USA. Association for Computational Linguistics.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2024b. Cofca: A step-wise counterfactual multi-hop qa benchmark. *Preprint*, arXiv:2402.11924.

Junda Wu, Xintong Li, Ruoyu Wang, Yu Xia, Yuxin Xiong, Jianing Wang, Tong Yu, Xiang Chen, Branislav Kveton, Lina Yao, Jingbo Shang, and Julian McAuley. 2024c. Ocean: Offline chain-of-thought evaluation and alignment in large language models. *Preprint*, arXiv:2410.23703.

Yexin Wu, Zhuosheng Zhang, and Hai Zhao. 2024d. Mitigating misleading chain-of-thought reasoning with selective filtering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11325–11340, Torino, Italia. ELRA and ICCL.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. *Preprint*, arXiv:2404.05692.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P. Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *Preprint*, arXiv:2405.00451.

Zhengnan Xie, Sebastian Thiem, Jaycie Martin, Elizabeth Wainwright, Steven Marmorstein, and Peter Jansen. 2020. WorldTree v2: A corpus of science-domain structured explanations and inference patterns supporting multi-hop inference. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5456–5473, Marseille, France. European Language Resources Association.

Wei Xiong, Hanning Zhang, Nan Jiang, and Tong Zhang. 2024. An implementation of generative prm. `https://github.com/RLHFlow/RLHF-Reward-Modeling`.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search. *Preprint*, arXiv:2205.12443.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. 2023. LeanDojo: Theorem proving with retrieval-augmented language models. In *Neural Information Processing Systems (NeurIPS)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Peiran Yao and Denilson Barbosa. 2024. Accurate and nuanced open-QA evaluation through textual entailment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2575–2587, Bangkok, Thailand. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024a. Free process rewards without process labels. *Preprint*, arXiv:2412.01981.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024b. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*.

Zhongshen Zeng, Pengguang Chen, Shu Liu, Haiyun Jiang, and Jiaya Jia. 2024a. Mr-gsm8k: A meta-reasoning benchmark for large language model evaluation. *Preprint*, arXiv:2312.17080.

Zhongshen Zeng, Yinhong Liu, Yingjia Wan, Jingyao Li, Pengguang Chen, Jianbo Dai, Yuxuan Yao, Rongwu Xu, Zehan Qi, Wanru Zhao, Linling Shen, Jianqiao Lu, Haochen Tan, Yukang Chen, Hao Zhang, Zhan Shi, Bailin Wang, Zhijiang Guo, and Jiaya Jia. 2024b. Mr-ben: A meta-reasoning benchmark for evaluating system-2 thinking in llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 119466–119546. Curran Associates, Inc.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Fengji Zhang, Bei Chen, Yue Zhang, Jacky Keung, Jin Liu, Daoguang Zan, Yi Mao, Jian-Guang Lou, and Weizhu Chen. 2023a. RepoCoder: Repository-level code completion through iterative retrieval and generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2484, Singapore. Association for Computational Linguistics.

Jiaxin Zhang, Zhong-Zhi Li, Ming-Liang Zhang, Fei Yin, Cheng-Lin Liu, and Yashar Moshfeghi. 2024a. GeoEval: Benchmark for evaluating LLMs and multimodal models on geometry problem-solving. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1258–1276, Bangkok, Thailand. Association for Computational Linguistics.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024b. Generative verifiers: Reward modeling as next-token prediction. *Preprint*, arXiv:2408.15240.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023b. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore. Association for Computational Linguistics.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023c. Enhancing uncertainty-based hallucination detection with stronger focus. *Preprint*, arXiv:2311.13230.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024c. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Preprint*, arXiv:2406.09136.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. *Preprint*, arXiv:2501.07301.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *Preprint*, arXiv:2411.14405.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024. Processbench: Identifying process errors in mathematical reasoning. *Preprint*, arXiv:2412.06559.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2021. Ar-lsat: Investigating analytical reasoning of text. *Preprint*, arXiv:2104.06598.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. 2024. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *Preprint*, arXiv:2410.23856.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024a. FanOutQA: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37, Bangkok, Thailand. Association for Computational Linguistics.

Tinghui Zhu, Kai Zhang, Jian Xie, and Yu Su. 2024b. Deductive beam search: Decoding deducible rationale for chain-of-thought reasoning. *Preprint*, arXiv:2401.17686.

A Tasks

This section aims to describe different reasoning tasks and datasets in more detail.

A.1 Multi-hop Question Answering

This section focuses on the metrics proposed for evaluating the reasoning traces for multi-hop question answering (MHQA) tasks. MHQA is often divided into two subcategories, **factual reasoning** and **commonsense reasoning**.

Inference in factual MHQAs is finding the sequence of *bridging entities* that leads to the final answer (Yang et al., 2018; Talmor and Berant, 2018; Kwiatkowski et al., 2019). For example, to solve a factual MHQA question *"The Argentine PGA Championship record holder has won how many tournaments worldwide?"*, one must first find who (*bridging entity*) is the Argentine PGA championship record holder and determine how many tournaments he has won worldwide.

In contrast, an inference step in commonsense MHQAs (Clark et al., 2018; Mihaylov et al., 2018; Talmor et al., 2019; Bisk et al., 2019; Geva et al., 2021; Trivedi et al., 2022) can require information that is not present in the provided facts. The form of such commonsense knowledge can be diverse, ranging from well-known facts (*Paris is in France.*) to logical rules (*If A was born after B was dead, they have never met each other*).

LLMs are known to achieve strong performance in challenging datasets such as ARC-Challenge and PIQA (OpenAI, 2024a; Anil et al., 2023), sometimes exceeding human performance. However, despite the high performance, multiple studies report that even modern LLMs like GPT-4 are vulnerable to errors, such as failing to correctly adhere to long evidence (Zhu et al., 2024a), leveraging shortcuts (Schnitzler et al., 2024), or ignoring temporal relation between events (Li et al., 2024a). Therefore, identifying and categorizing mistakes made by LLMs can still be considered relevant tasks.

A.2 Symbolic Reasoning

Due to the improvement of LLMs' reasoning ability since the discovery of Chain-of-thought prompting (Wei et al., 2022b; Kojima et al., 2022), step-by-step reasoning has proven effective in symbolic reasoning tasks[5] such as **mathematical reasoning**,

[5]While symbolic reasoning may strictly refer to *algorithmic reasoning* (Wei et al., 2022b), we adopt the broader sense including math and logical reasoning that can be readily expressed in symbols (equation, logic) (Sprague et al., 2024).

logical reasoning, and **algorithmic reasoning**.

Arithmetic reasoning, where the model has to predict the correct answer from arithmetic word problems, is the most recognized variant of math reasoning. Popular benchmarks include MathQA (Amini et al., 2019) and GSM8k (Cobbe et al., 2021), which provide long, diverse natural language queries. Game of 24 (Yao et al., 2023) and Mathador (Kurtic et al., 2024) ask to combine given numbers and arithmetic operations to generate the target number, requiring thorough exploration and backtracking in the exponential solution space.

The rapid saturation of LLMs in arithmetic word problems facilitated more challenging **mathematical reasoning** benchmarks from olympiad/graduate-level problems, covering fields like calculus, probability and statistics, geometry, number theory, and more (He et al., 2024a; Gao et al., 2024b; Glazer et al., 2024; Zhang et al., 2024a). Recent reasoning-focused (*a.k.a.* slow-thinking) LLMs (OpenAI, 2024b; Qwen-Team, 2024; DeepSeek-AI, 2025) achieve unprecedented performance in these benchmarks by generating long reasoning traces with self-verification and correction.

Deductive logical reasoning (Tafjord et al., 2021; Tian et al., 2021; Saparov and He, 2023; Han et al., 2024a) mainly focuses on logical deduction, where repeatedly applying the provided rules to facts will reach the correct answer. **Constraint-based reasoning** (Zhong et al., 2021; Tyagi et al., 2024) is a variant of deductive reasoning where one must find the solution that suffices the provided initial constraints (also referred to as *grid puzzle*). These datasets have an exponentially sized solution space that significantly reduces the LLM's reasoning performance in plain Chain-of-thought setting (Kang et al., 2024).

Finally, **algorithmic (symbolic) reasoning** tasks include manipulating strings and data structures, such as concatenating the last letters of the given words (Wei et al., 2022b) or completing the incomplete Dyck language. BIG-Bench-Hard (BBH; Suzgun et al. (2022)) and NPHardEval (Fan et al., 2024) includes 11 and 9 algorithmic reasoning tasks, respectively, which is challenging for even modern LLMs like GPT-4 and PaLM-540B.

A.3 Uncovered tasks

Science reasoning tasks lie between factual/commonsense reasoning tasks and symbolic reasoning tasks, as they often require addressing very complicated facts and performing precise math/logical reasoning (Rein et al., 2024; He et al., 2024a). The most popular benchmark in this field, GPQA-Diamond (Rein et al., 2024), contains 546 questions from physics, chemistry, and biology where human experts only get 65% of the problem correct.

Programming/coding is closely related to algorithmic reasoning tasks. Popular benchmarks regarding programming include *competitive coding* where one has to solve an algorithm problem given in natural language and test codes (Chen et al., 2021; Li et al., 2022), and *practical coding* that covers tasks of software engineers and developers (Zhang et al., 2023a; Jimenez et al., 2024; Chan et al., 2024). While writing a correct program requires reasoning ability, coding differs from other reasoning tasks in various aspects including: (1) there is a strict syntax requirement for code, and (2) the result is evaluated by the execution result, not the final answer. These constraints lead to several issues when (1) segmenting the trace (code) into steps, or (2) applying metrics that require explicitly stated answers, *i.e.* $\mathcal{V}$-information.

B Resources

This section enumerates useful resources containing stepwise annotation. These datasets can be used to train an evaluator or perform meta-evaluation on different metrics.

B.1 Factual/Commonsense reasoning

For meta-evaluating metrics in factual/commonsense reasoning, human annotations on LLM-generated outputs are provided by ROSCOE (Golovneva et al., 2023a), REVEAL (Jacovi et al., 2024), and MR-Ben (Zeng et al., 2024b) (MMLU portion).

B.2 Symbolic Reasoning

Training data for *validity* evaluators. The most popular validity dataset used for training PRMs is PRM800k (Lightman et al., 2024), which contains 800k human-anntoated stepwise labels (75k reasoning traces) in MATH (Hendrycks et al., 2021) dataset. It classifies each step into three labels, *positive, neutral, and negative*, where *negative* denotes a clearly incorrect step and *neutral* is used to defer the annotator's uncertainty in borderline cases. Other than PRM800k, MATH-Minos (Gao et al., 2024a) provides LLM-generated validity judgments for 440k reasoning traces.

Dataset	Train	Eval	Base task	Criteria	# Trace	Human
ROSCOE (Golovneva et al., 2023b)		●	GSM8k, DROP, eSNLI, COSMOS-QA, SemEval-2018 Task11	`(GV)U`	1.0k	●
REVEAL (Jacovi et al., 2024)		●	StrategyQA, MuSiQue, Sports, Fermi	`G(VC)`	3.4k	●
PRM800k (Lightman et al., 2024)	●	●	MATH	`V`	75k	●
MATH-Minos (Gao et al., 2024a)	●		GSM8k, MATH	`V`	440k	×
SCDPO (Lu et al., 2024)	●		GSM8k, MATH	`U`	30k	×
MR-GSM8k (Zeng et al., 2024a)		●	GSM8k	`V`	3.0k	●
MR-Ben (Zeng et al., 2024b)		●	MMLU (science), LogiQA, MHPP (coding)	`V`	6.0k	●
MR-MATH (Xia et al., 2025)		●	MATH	`V`	0.1k	●
BIG-Bench-Mistake (Tyen et al., 2024)		●	BIG-Bench (algorithmic)	`(VC)U`	2.2k	●
ProcessBench (Zheng et al., 2024)		●	GSM8k, MATH, Olympiad-Bench, Omni-MATH	`V`	3.4k	●
PRMBench (Song et al., 2025)		●	MATH	`VCU`	6.2k	▲
Math-Shepherd (Wang et al., 2024b)	●		GSM8k, MATH	`U`	440k	×

Table 2: List of PRM training data and meta-evaluation benchmarks with step-wise annotation. **Train/Eval** columns denote if the dataset is used for training or meta-evaluation. **Base task** indicates what tasks are used to sample the reasoning trace. **Criteria** column shows the criteria used to annotate the data classified according to Section 3, where `GVCU` stands for groundedness, validity, coherence, and utility, respectively. Parentheses indicate that the criteria group is not explicitly distinguished in the labels. **Human** column indicates human annotation, where ● ▲ × denotes full human annotation, automatic annotation with human verification, and no human intervention, respectively.

Meta-evaluating *validity* evaluators. There are multiple validity meta-evaluation benchmarks that incorporate human evaluation. PRM800K (Lightman et al., 2024), MR-GSM8k (Zeng et al., 2024a), MR-Ben (Zeng et al., 2024b), MR-MATH (Xia et al., 2025), BIG-Bench-Mistake (Tyen et al., 2024), ProcessBench (Zheng et al., 2024), and PRMBench (Song et al., 2025). PRM800k, BIG-Bench-Mistake, and PRMBench formulate the task as stepwise classification, where one has to evaluate each step logically correct or not. In contrast, ProcessBench and MR-* series are set to identify the index of the first erroneous step in the reasoning trace.

Training data for *utility* evaluators. Training data for *utility* evaluators. The most popular option is Math-Shepherd (Wang et al., 2024b), which includes 445k reasoning traces with labels assigned by MCTS. A step's label is positive if any of the $N = 8$ rollouts starting from the step leads to a correct answer, and negative otherwise. Also, Step-Controlled DPO (Lu et al., 2024) provides a large set of correct and incorrect reasoning traces, where incorrect ones are obtained by slowly increasing the LLM's temperature.

Meta-evaluating *utility* evaluators. The standard approach for utility meta-evaluation in symbolic reasoning is applying **Best-of-N (BoN)** decoding on challenging math reasoning datasets (Wang et al., 2024b; Cui et al., 2024; Zhang et al., 2025). In this setting the evaluator should choose the best trace among N sampled candidates, and the answer accuracy is determined from the selected one. A slight variant, **weighted voting** (Yuan et al., 2024a), decides the final answer based on the sum of evaluation scores instead of choosing the one with the highest score. In both settings, the upper bound of utility evaluators' performance is `pass@N` score, which counts when at least one from N traces has a correct answer.

C Metrics

Criterion	Implementation	Works
Groundedness	Rule-based	PrOntoQA† (Saparov and He, 2023), Nguyen et al. (2024)
	Uncertainty	SynCheck (Wu et al., 2024a), Qiu et al. (2024), Zhang et al. (2023c), Semantic entropy probes (Farquhar et al., 2024; Kossen et al., 2024)
	Cross-encoders	ROSCOE-LI (Golovneva et al., 2023b), ReCEval (Prasad et al., 2023), DBS (Zhu et al., 2024b), SynCheck (Wu et al., 2024a), *As a baseline* (Jacovi et al., 2024)
	PRMs	*As a baseline* (Song et al., 2025)
	Critic models	RAGTruth (Niu et al., 2024), OCEAN (Wu et al., 2024c), F^2-Verification (Wang et al., 2024a), *As a baseline* (Ling et al., 2023; Jacovi et al., 2024; Song et al., 2025, *inter alia.*)
Validity	Rule-based	PrOntoQA† (Saparov and He, 2023), Nguyen et al. (2024), DiVeRSe (Li et al., 2023)
	$\mathcal{V}$-information	ReCEval (Prasad et al., 2023)
	Cross-encoders	ROSCOE-LI (Golovneva et al., 2023a), ReCEval (Prasad et al., 2023), DBS (Zhu et al., 2024b), *As a baseline* (Jacovi et al., 2024)
	PRMs	PRM800K (Lightman et al., 2024), MATH-Minos (Gao et al., 2024a), ReasonEval (Xia et al., 2025), Qwen-PRM (Zhang et al., 2025), *As a baseline* (Zheng et al., 2024; Zeng et al., 2024b; Xia et al., 2025; Song et al., 2025, *inter alia.*)
	Critic models	F^2-Verification (Wang et al., 2024a), *As a baseline* (Ling et al., 2023; Jacovi et al., 2024; Zheng et al., 2024; Song et al., 2025, *inter alia.*)
	Generative verifiers	CLoud (Ankner et al., 2024), Generative verifier (Zhang et al., 2024b)
Coherence	Rule-based	PrOntoQA† (Saparov and He, 2023), Nguyen et al. (2024)
	Cross-encoders	ROSCOE-LI* (Golovneva et al., 2023a), DiVeRSe (Li et al., 2023), DBS (Zhu et al., 2024b)
	PRMs	*As a baseline* (Wang et al., 2024b)
	Critic models	Verify-CoT (Ling et al., 2023), *As a baseline* (Song et al., 2025)
Utility	Rule-based	PrOntoQA† (Saparov and He, 2023), DiVeRSe (Li et al., 2023), Nguyen et al. (2024)
	Uncertainty	Chain-of-probe (Wang et al., 2024d)
	$\mathcal{V}$-information	REV (Chen et al., 2023), ReCEval (Prasad et al., 2023), *As a baseline* (Yao and Barbosa, 2024)
	Cross-encoders	DBS (Zhu et al., 2024b)
	PRMs	Math-Shepherd (Wang et al., 2024b), RLHFlow-PRM (Xiong et al., 2024), Skywork-o1-Open-PRM (o1 Team, 2024), Eurus-PRM (Yuan et al., 2024b), Qwen-PRM (Zhang et al., 2025)
	Critic models	Tree-of-thoughts (Yao et al., 2023), CPO (Zhang et al., 2024c), CriticBench (Lin et al., 2024), *As a baseline* (Song et al., 2025)
	LLM-as-value-function	GenRM (Mahan et al., 2024), Step-DPO (Lai et al., 2024), MCTS-DPO (Xie et al., 2024), IRPO (Pang et al., 2024) Tree-PLV (He et al., 2024b), Step-Controlled DPO (Lu et al., 2024), PRIME (Cui et al., 2025)

Table 3: Metrics for step-by-step reasoning evaluation, sorted by criteria and implementation. If a work falls into multiple categories because it ensembles different metrics or proposes/tests multiple implementations, it appears as duplicate entities in the table. For the construction of Table 1, we do not count works with the following marks: **While not explicitly claimed, the training instances might include errors about the criterion.* †*While the work proposes a metric, the implementation is not transferable to other datasets.*