

BOOSTING PROCESS-CORRECT CoT REASONING BY MODELING SOLVABILITY OF MULTIPLE-CHOICE QA

Anonymous authors

Paper under double-blind review

ABSTRACT

Reasoning quality in large language models depends not only on producing correct answers but also on generating valid intermediate steps. We study this through multiple-choice question answering (MCQA), which provides a controlled setting with fixed answer options. Our analysis shows that when questions are effectively unsolvable for a model, spurious chains of thought (CoTs) are more likely to appear, leading to false positives. By estimating the solvability of each question, we uncover an intermediate regime where learning is most effective. Building on this insight, we adapt outcome-supervised reward models and reinforcement learning with group-relative advantage to incorporate solvability into their objectives. Across experiments on math and multimodal datasets, these modifications consistently yield higher rates of process-correct reasoning and, in reinforcement learning, improved answer accuracy as well. Our results highlight solvability as a key factor for reducing hallucinations and increasing reliability in CoT reasoning.

1 INTRODUCTION

In many applications of CoT reasoning, the generated thought process is as important as the final answer. While some tasks provide gold-standard reasoning chains that can effectively be used for supervised training (Nye et al., 2021; Dziri et al., 2023; Hochlehnert et al., 2025), most datasets lack such annotations. For these cases, correct reasoning has to be incentivized by rewards on correct final answers (Wen et al., 2025). It is known that CoTs can lead to the correct answer, despite an incorrect explanation. Grattafiori et al. (2024) note that this often occurs for questions where only a small fraction of the generated answers is correct. In this work, we investigate this observation in controlled experiments on multiple datasets. To avoid confounding factors of noisy answer extraction and matching, we focus on multiple-choice question answering. This format is popular for evaluating models and widely used training sets like NuminaMath (LI et al., 2024) contain a large fraction of multiple-choice questions. The fixed number of answer options also allows us to explicitly model the solvability of a question. We find that unsolvable questions promote false positive CoTs. Additionally, in a controlled finetuning experiment we show that there is a sweet spot of questions for which neither a small nor high fraction of CoTs lead to the correct answer. We make use of these findings by modifying the objective function of an outcome-based reward model (ORM) and by adjusting the advantage calculation of group relative reinforcement learning (RL). The proposed modifications lead to more process-correct CoTs and additionally, in the case of RL, to improved answer accuracy. Furthermore, following the argumentation of Kalai & Vempala (2024); Kalai et al. (2025) according to which LLMs are optimized to guess when uncertain, such hallucinations can be mitigated by modeling solvability in the learning objective.

2 BACKGROUND

In chain-of-thought (CoT) reasoning, a model π_θ is presented with a question q_i and prompted to generate an output o_{ij} consisting of a thought process t_{ij} and final answer \hat{y}_{ij} :

$$o_{ij} \sim \pi_\theta(\cdot | q_i), \quad \text{where } o_{ij} = (t_{ij}, \hat{y}_{ij}). \quad (1)$$

To ensure diverse outputs when sampling multiple CoTs per question, the token logits are divided by a positive temperature value. In this work, we use temperature 1.0 for all experiments and reported

results. Given the ground-truth answer y_i , the correctness of the generated answer is determined by a binary scoring function $\mathbb{1}[y_i = \hat{y}_{ij}]$ that equals 1 if $y_i = \hat{y}_{ij}$ and 0 otherwise. Because matching of open-ended answers can be ambiguous, multiple-choice question answering (MCQA) is a popular format. Each question includes a letter-indexed list of predefined answer choices c_i with exactly one correct choice, and scoring reduces to exact letter matching.

Metrics Given a dataset \mathbb{D} with question-answer pairs $(q_i, y_i) \in \mathbb{D}$, the performance of CoT reasoning is commonly measured by answer accuracy (A-Acc). The generated answer \hat{y}_{ij} is compared with the ground truth answer y_i and the binary score is averaged across questions and samples:

$$\text{A-Acc} := \frac{1}{|\mathbb{D}|G} \sum_{i=1}^{|\mathbb{D}|} \sum_{j=1}^G \mathbb{1}[y_i = \hat{y}_{ij}]. \quad (2)$$

A CoT consists of the thought process and the final answer. The correctness of the latter can be determined by comparing it to the ground-truth answer. Because there does not exist *the one* ground-truth thought process, we employ an LLM to judge its correctness. Although this is common practice, it has to be handled with care (He et al., 2024; Hao et al., 2024; Bavaresco et al., 2025). As such, we conduct an extensive meta-evaluation of the judge by reporting correlation with human judgments, measuring performance on a synthetic dataset, and manually evaluating a subset of the judgments (Appendix D). In addition, we release all outputs verbatim to facilitate future comparisons with our work. Formally, the judge \mathcal{J}_{LLM} receives the question q_i , a thought process t_{ij} , and the ground-truth answer y_i as input, and returns the binary judgment:

$$\mathcal{J}_{LLM} : (q_i, t_{ij}, y_i) \rightarrow \{0, 1\}. \quad (3)$$

We then calculate the process accuracy (P-Acc) as:

$$\text{P-Acc} := \frac{1}{|\mathbb{D}_{AC}|} \sum_{q_i, t_{ij}, y_i \in \mathbb{D}_{AC}} \mathcal{J}_{LLM}(q_i, t_{ij}, y_i). \quad (4)$$

In this work, we report process accuracy solely on the subset of answer-correct CoTs, denoted as \mathbb{D}_{AC} . A CoT that is answer-correct but process-incorrect is referred to as *false positive*.

Outcome-Supervised Reward Model An outcome-supervised reward model (ORM) is used to predict the correctness of a generated answer without access to the ground-truth (Cobbe et al., 2021). It is denoted as π_ϕ and its training objective is to minimize the binary cross entropy loss:

$$\mathcal{L}_{ij}^{\text{BCE}}(\pi_\phi) = -z_{ij} \log \pi_\phi(h_{ij}) - (1 - z_{ij}) \log(1 - \pi_\phi(h_{ij})) \quad (5)$$

where $z_{ij} = \mathbb{1}[y_i = \hat{y}_{ij}]$ is the label. The input h_{ij} is the representation of the question and the sampled output (q_i, o_{ij}) , e.g., the raw text or the LLM’s last hidden state during generation. The ORM can then be used to rerank outputs at test-time or as reward estimator in reinforcement learning.

Reinforcement Learning with Group Relative Advantage Recently, reinforcement learning with estimation of group relative advantage has gained renewed traction, especially in domains with verifiable rewards (Kool et al., 2019; Shao et al., 2024; DeepSeek-AI et al., 2025). The approach eliminates the complexity of training a reward model and the accompanying problems such as reward hacking or data bias. Instead, the advantage or value of an action o_{ij} is determined relative to other samples for the same input. This means we sample multiple outputs per question and the reward for each output is computed as $r_{ij} = \mathbb{1}[y_i = \hat{y}_{ij}]$, where \hat{y}_{ij} is *null* if no answer can be extracted from o_{ij} , e.g. due to incorrect answer format. The advantage is then calculated relative to the rewards of the other samples. Specifically, the formulations of the GRPO (Shao et al., 2024) and DrGRPO (Liu et al., 2025) variants are:

$$A_{ij}^{\text{GRPO}} = \frac{1}{\sigma(\mathbf{r}_i)}(r_{ij} - \frac{1}{G} \sum_{k=1}^G r_{ik}) \quad \text{and} \quad A_{ij}^{\text{DrGRPO}} = r_{ij} - \frac{1}{G} \sum_{k=1}^G r_{ik}, \quad (6)$$

respectively, where \mathbf{r}_i is the reward vector and $\sigma(\cdot)$ returns the standard deviation. The policy gradient for a single question, simplified here without the standard PPO clipping term, becomes:

$$\nabla_\theta J(\theta) \approx \frac{1}{G} \sum_{j=1}^G A_{ij} \nabla_\theta \log \pi_\theta(o_{ij} | q_i), \quad (7)$$

where the advantage A_{ij} of a sample o_{ij} is computed using GRPO or DrGRPO as defined above.

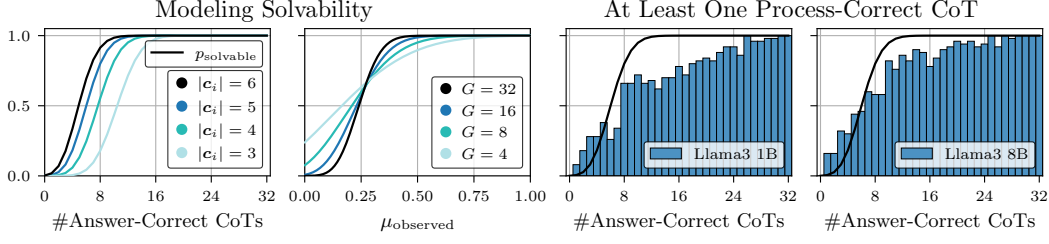


Figure 1: **Modeling Solvability:** The probability that a question is solvable by a given model, as defined by Equation 11. (Left) Varying number of answer options for the multiple-choice question. (Right) Varying number of sampled CoTs per question. **At Least One Process-Correct CoT:** Fraction of questions where at least one of the 32 generated CoTs is process-correct. Questions are from the AQuA dataset (five choices) and CoTs are sampled with Llama3 1B (left) and Llama3 8B (right).

3 SOLVABILITY OF MULTIPLE-CHOICE QUESTIONS

We model the binary outcome of whether a CoT o_{ij} , sampled from model π_θ , correctly answers a question q_i as a Bernoulli random variable:

$$\mathbb{E}_{o_{ij} \sim \pi_\theta(\cdot|q_i)} [\mathbb{1}[y_i = \hat{y}_{ij}]] = \mu_{\text{true}}^\theta(q_i). \quad (8)$$

The true parameter $\mu_{\text{true}}^\theta(q_i)$ is unobservable, but can be estimated by sampling G outputs:

$$\mu_{\text{observed}}^\theta(q_i) = \frac{1}{G} \sum_{j=1}^G \mathbb{1}[y_i = \hat{y}_{ij}]. \quad (9)$$

Using a uniform prior $\text{Beta}(1, 1)$ and the observed success rate $\mu_{\text{observed}}^\theta(q_i)$ as likelihood, the posterior distribution for $\mu_{\text{true}}^\theta(q_i)$ is given by $\text{Beta}(\alpha_i, \beta_i)$ with the parameters:

$$\alpha_i = 1 + G\mu_{\text{observed}}^\theta(q_i) \quad \text{and} \quad \beta_i = 1 + G(1 - \mu_{\text{observed}}^\theta(q_i)), \quad (10)$$

which represent the success and failure counts, respectively. We define a question as solvable by the model if the model’s true performance exceeds random guessing: $\mu_{\text{true}}^\theta(q_i) > \mu_{\text{random}}(q_i)$. The random guessing baseline for a multiple-choice question is given by: $\mu_{\text{random}}(q_i) = \frac{1}{|c_i|}$. Using this information, we compute the probability that a question is solvable for the model as the survival function of the Beta distribution:

$$p_{\text{solvable}}^\theta(q_i) = p(\mu_{\text{true}}^\theta(q_i) > \mu_{\text{random}}(q_i)) = \int_{\mu_{\text{random}}(q_i)}^1 \text{Beta}(\mu; \alpha_i, \beta_i) d\mu. \quad (11)$$

The left section of Figure 1 illustrates $p_{\text{solvable}}^\theta(q_i)$ for varying number of answer choices $|c_i|$ and varying number of samples G . When only a small fraction of CoTs yield the correct answer, the model’s probability of solving the question approaches zero. As the number of answer-correct CoTs increases, solvability rises exponentially before converging to unity. Both the onset and the inflection point depend on the number of answer choices. The more answer choices a question offers, the smaller the proportion of correct CoTs that is required to achieve solvability. The steepness of the increase depends on the number of sampled CoTs — more samples provide a clearer distinction between solvable and unsolvable questions.

3.1 SOLVABILITY AND PROCESS-CORRECTNESS

Intuitively, if a question is not solvable for a model, the model should not be able to generate a CoT with correct thought process. We empirically verify this intuition in Figure 1 (right section). There, the questions in the math reasoning dataset AQuA (Ling et al., 2017) are categorized by the number of answer-correct CoTs, generated by Llama3 1B and Llama3 8B (Grattafiori et al., 2024). The $p_{\text{solvable}}^\theta(q_i)$ line closely follows the empirical data in the bar chart, showing it is a good predictor of whether the model is able to generate a correct thought process for a given question. By incorporating this probability into the training of an outcome reward model (Section 4) and advantage calculation of reinforcement learning (Section 5), we expect to boost the ability to identify and generate process-correct CoTs, respectively.

Table 1: **Process-Accuracy (P-Acc)** using different methods of CoT scoring. The task is to score multiple candidate CoTs that all lead to the correct answer. The highest scoring CoT is then evaluated for process-correctness. This is done for 200 questions of the three multiple-choice QA datasets. Oracle gives the upper bound because not every candidate set contains a process-correct CoT. Outcome-supervised reward models (ORM) are trained with three different random seeds and mean \pm std is reported. Nominal best values are bold.

	Llama3 1B			Llama3 8B		
	AQuA	MATH	GSM8K	AQuA	MATH	GSM8K
Oracle	79.5	80.0	93.5	96.0	92.0	98.0
Random	47.0	45.7	66.0	81.5	63.3	90.2
Shortest	47.0	51.5	65.0	87.0	75.0	94.0
Longest	26.0	27.0	47.5	53.0	36.0	61.5
CoPS (Wang et al., 2025)	52.0	54.5	69.5	61.5	49.5	72.5
Faithfulness (Paul et al., 2024)	37.5	39.5	58.0	71.0	56.0	87.0
Answer Confidence	57.5	48.5	76.5	84.0	73.5	92.0
ORM	67.3 \pm 0.8	64.5 \pm 1.5	87.8 \pm 1.0	90.3 \pm 0.2	81.0 \pm 0.8	95.7 \pm 0.2
MCQ-ORM (ours)	70.0 \pm 0.7	65.7 \pm 0.2	88.7 \pm 0.5	92.0 \pm 0.4	83.5 \pm 0.4	96.2 \pm 0.5

4 BOOSTING PROCESS-CORRECTNESS AT TEST-TIME

A common technique to improve answer accuracy at test-time is to sample multiple CoTs for a question and select the final answer by majority vote (Wang et al., 2023b). This means that there are multiple candidate CoTs that lead to the majority-voted answer. In this section, we discuss the task of selecting the candidate CoT that is most likely process-correct. Uesato et al. (2023) use the score of an ORM to select the most promising CoT among the candidates. They show that this improves average process-correctness in comparison to random selection. Their ORM is trained with binary outcome labels where answer-correct CoTs have label 1, and 0 otherwise. Instead, we incorporate the probability that a question is solvable (Equation 11) into the ORM objective (Equation 5):

$$z_{ij} = \begin{cases} p_{\text{solvable}}^{\theta}(q_i), & \text{if } \hat{y}_{ij} = y_i \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

This gives lower weight to CoTs that are likely *false positive* and therefore should receive a lower score when ranking the candidate CoTs during test-time. We call this modification MCQ-ORM to reflect the incorporation of solvability of multiple-choice questions.

4.1 EXPERIMENTS

We train the proposed MCQ-ORM and estimate its accuracy in selecting a process-correct CoT among multiple answer-correct CoTs. We compare it to the unmodified ORM and other baselines. As base models we use Llama3 with 1B and 8B parameters. The training and development set for the reward models are sourced from the 97.5k training questions of AQuA. We report process accuracy for all methods on three math reasoning datasets for both base models.

Reward Model Training Given a base model, we sample 32 CoTs for each of the 97.5k training questions, resulting in 3M training and 32k development instances for the reward model. A training instance consists of the base model’s last hidden state as input and the appropriate outcome-based label (see Equation 5 for the unmodified ORM and Equation 12 for our proposed MCQ-ORM). The architecture is a feed-forward neural network with two hidden layers and sigmoid activations. Additional hyperparameters are optimized individually for each reward model (Appendix C). Both ORM variants are trained with three different random seeds, and early stopping is based on cross-entropy loss of the development set.

Evaluation In addition to AQuA (five choices), we also report results on the MATH (Hendrycks et al., 2021) and GSM8K (Cobbe et al., 2021) datasets. Both are modified to follow the MCQA format with four choices and six choices, respectively (Zhang et al., 2024). For each test question,

we sample 32 CoTs and determine the predicted answer by majority vote. The subset of CoTs that lead to the majority answer is the candidate set. The task is to select one CoT among the candidates that is most likely process-correct. Each method is evaluated on the same 200 questions from each dataset ranging from 6 to 32 candidate CoTs. Because it is not guaranteed that at least one candidate is process-correct, we also report oracle results as an upper bound.

4.2 RESULTS

The results in Table 1 show that both base models are generally capable of generating process-correct CoTs. The large gap between the random baseline and the oracle shows substantial room to test the capabilities of the considered selection methods. The CoPS (Wang et al., 2025) and faithfulness (Paul et al., 2024) baselines are making use of early answer probing. After each reasoning step, the model is forced to decode the correct answer letter and its token probability is recorded in a vector. This early answer probability vector is then used to draw conclusions about the reasoning process. CoPS estimates the quality of a CoT by considering the average probability of early answers and their increase over time. The CoT faithfulness metric was developed to measure the alignment of the model’s internal with its external textual reasoning. It is defined as the area over the curve of the early answering probability vector. We can see that both of these metrics are outperformed by a simple baseline that ranks CoTs by final answer confidence, i.e., the probability assigned to the correct answer letter. Using the score of a reward model to select the best CoT largely outperforms the aforementioned baselines. Our proposed MCQ-ORM that takes the solvability of a question into account consistently outperforms the unmodified ORM across considered datasets and base models. Although the effect size is small, a random permutation test (Appendix B) shows that the results are overall significant.

5 REINFORCEMENT LEARNING WITH ADJUSTED ADVANTAGE

We start with an analysis of advantage values calculated by GRPO and DrGRPO. The plots on the left and middle of Figure 2 show that a sample o_{ij} with positive reward $r_{ij} = 1$ gets the highest individual advantage if all other samples in the group received a negative reward. Comparing this to the right section of Figure 1, we see that these samples correspond to CoTs that are most likely process-incorrect. To further investigate the impact of this advantage shape, we conduct a controlled experiment where we estimate the learning potential of question-CoT pairs, depending on the number of answer-correct CoTs in the group.

5.1 ESTIMATING LEARNING POTENTIAL

We sample 32 CoTs for each question in the respective training set. The questions are then categorized into buckets based on the number of answer-correct CoTs in the group. We then randomly select a subset of the questions in a bucket. For each of the questions in the subset, we randomly select exactly one of its answer-correct CoTs. Formally, a finetuning dataset for bucket b is:

$$\mathbb{D}_b^{FT} = \{(q_i, o_{ij}) \mid q_i \sim Q_b, j \sim \{k : r_{ik} = 1\}\}, \text{ where } Q_b = \{q_i \mid \sum_{k=1}^G r_{ik} = b\}. \quad (13)$$

For each bucket, such a dataset with 2k instances is sampled and used to finetune the base model. We then evaluate the finetuned models on the development set of the respective dataset and measure the improvement in answer-accuracy over the base model. Figure 3 shows the results for AQuA (Ling et al., 2017), MedMCQA (Pal et al., 2022) and SocialQA (Sap et al., 2019) using Llama2 7B (Touvron et al., 2023). Each experiment is repeated five times with different random seeds and the mean is depicted. The seed affects the subset selection and the order of training batches. Buckets with less than 2k questions are merged with their neighbor bucket. For all datasets, the observed distribution of accuracy improvement is left-skewed, with a linear decrease to the right and a steep increase on the left. The position of the distribution mode varies depending on the dataset. In the following, we derive a simple model that describes the observed accuracy improvement of question-CoT pairs based on bucket membership and number of answer choices.

Given a question and sampled CoTs, we seek to model the learning potential (LP) of a pair (q_i, o_{ij}) .

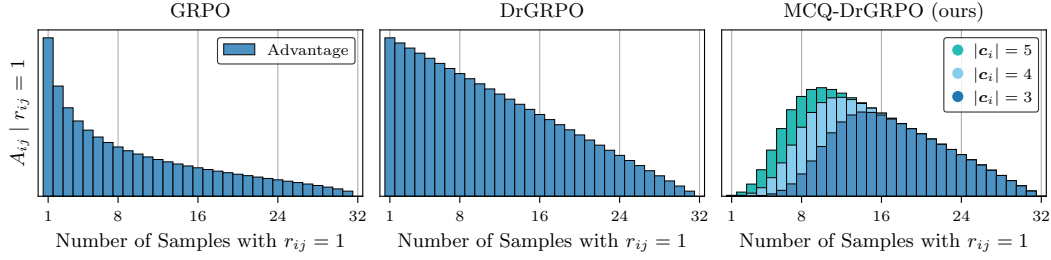


Figure 2: Advantage values of a single CoT with positive reward. 32 CoTs are sampled for each question and the x-axis denotes the number of answer-correct (positive reward) CoTs in a group. MCQ-DrGRPO down-weights CoTs that are generated for unsolvable questions. The probability that a multiple-choice question is unsolvable for the model depends on the number of choices $|c_i|$. The values on the y-axis are omitted to allow visual comparison across methods. During training the relative differences between groups are important.

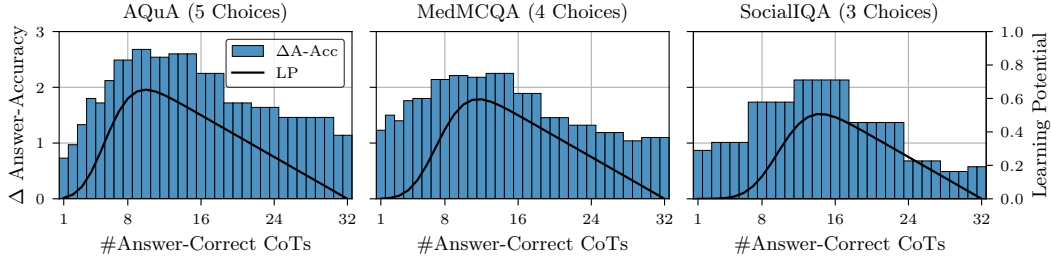


Figure 3: We sample 32 CoTs for each question in the respective training set. Questions are then categorized into buckets based on the number of answer-correct CoTs. We randomly sample questions from each bucket and pair them with exactly one of their answer-correct CoTs. We finetune the base model on these 2k instances and report the increase in answer accuracy over the base model on a held out development set. Experiments are repeated five times with different random seeds. Learning potential (LP) predicts relative increase in answer accuracy based on bucket membership.

Questions that are trivially solved by the model offer minimal informational gain, as they lack novelty with respect to the model’s existing knowledge. In contrast, questions that the model fails to answer correctly contain maximal novel information. We formalize the probability that a question provides novel information as the fraction of incorrect answers:

$$p_{\text{novel}}^{\theta}(q_i) = \frac{1}{G} \sum_{j=1}^G \mathbb{1}[y_i \neq \hat{y}_{ij}]. \quad (14)$$

However, as seen in Figure 3, a counteracting mechanism limits the model to learn from overly novel inputs. Specifically, when a question exceeds the model’s current capabilities, the learning signal becomes noisy or cannot be effectively utilized. We capture this trade-off through the following formulation:

$$\text{LP}(q_i, o_{ij}) = p_{\text{novel}}^{\theta}(q_i) p_{\text{solvable}}^{\theta}(q_i). \quad (15)$$

The line in Figure 3 shows that this estimation of learning potential aligns well with the observed improvement in accuracy. We will use this finding to adjust the advantage calculation in order to prefer instances with high learning potential.

5.2 ADVANTAGE CALCULATION ADJUSTED BY SOLVABILITY

Using equality $\mathbb{1}[y_i = \hat{y}_{ij}] = 1 - \mathbb{1}[y_i \neq \hat{y}_{ij}]$, we can rearrange the DrGRPO advantage calculation (Equation 6) of a sample with positive reward to be equal to the novelty formulation in Equation 14. Applying our findings that there is a trade-off between novelty and solvability, we propose solvability-adjusted DrGRPO for multiple-choice questions:

$$A_{ij}^{\text{MCQ-DrGRPO}} = p_{\text{solvable}}^{\theta}(q_i) A_{ij}^{\text{DrGRPO}} \quad (16)$$

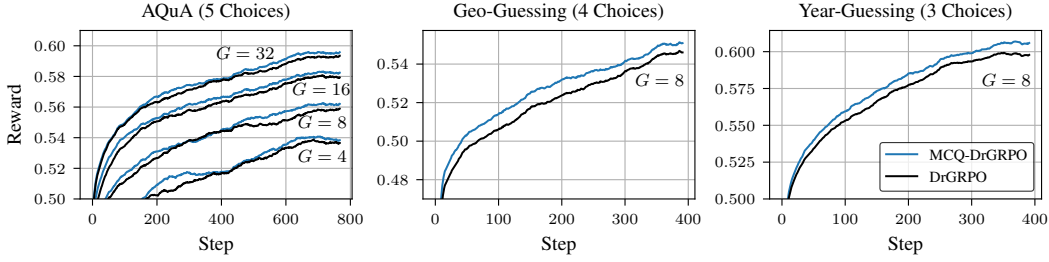


Figure 4: The average reward during RL training with DrGRPO and MCQ-DrGRPO. The math reasoning dataset AQuA is used to train Llama3 1B and geo/year-guessing datasets are used to train multimodal Aya 8B. First graph also shows ablations for different numbers of sampled CoTs per question. Each model is trained with three different random seeds.

The plot on the right in Figure 2 shows that this calculation focuses advantage mass on samples with high learning potential. CoTs sampled for unsolvable questions, and hence likely process-incorrect, are downweighted compared to GRPO and DrGRPO.

5.3 REINFORCEMENT LEARNING EXPERIMENTS

We perform reinforcement learning experiments using proximal policy optimization (Schulman et al., 2017) with group relative advantage estimation (Shao et al., 2024). Specifically, we compare the advantage estimation of DrGRPO (Equation 6) to our proposed MCQ-DrGRPO (Equation 16) that incorporates the solvability of a multiple-choice question into the advantage calculation. We use KL penalty and remove output length bias Liu et al. (2025). See Appendix C for more details of the implementation. Besides the answer accuracy (A-Acc) we also report the process accuracy (P-Acc) which is the average correctness of the thought processes that lead to a correct answer (Section 2).

Math Reasoning We use AQuA (five choices) as the training set and report evaluation metrics on two additional datasets. These are MATH and GSM8K, both modified to follow the MCQA format with four choices and six choices, respectively (Zhang et al., 2024). As base models, we use Qwen2.5 1.5B (Qwen et al., 2025) and Llama3 1B because it is not saturated on the considered datasets and the relatively small size allows us to conduct additional ablation experiments. We sample 32 CoTs per question during training.

Multimodal Reasoning Due to the lack of large-scale multimodal reasoning datasets that are *not* math-related, we construct two novel MCQA datasets. One asks for the geographic region (four choices) in which an image was taken and the other for the year (three choices) when it was taken. These tasks require the model to analyze different aspects of the image and combine it with general knowledge to draw a conclusion. Both datasets have 93k training instances, and a development and test set of size 3.5k each. See Appendix A for more details and download link. We use multimodal Aya 8B (Dash et al., 2025) as the base model and sample 8 CoTs per question during training.

5.4 RESULTS

The plots in Figure 4 show average rewards during training. Our proposed MCQ-DrGRPO achieves consistently higher rewards than the DrGRPO baseline. The experiments are repeated three times with different random seeds, affecting data ordering and token sampling. Table 2 and Table 3 show the process and answer accuracy, evaluated on three datasets. Both methods improve not only answer accuracy over the base model, but also process accuracy. This confirms recent findings that reinforcement learning with verifiable rewards (RLVR) implicitly optimizes the correctness of the thought process (Wen et al., 2025). Our proposed MCQ-DrGRPO consistently outperforms DrGRPO with larger effect sizes for process accuracy. This shows the effectiveness of downsizing advantage values of unsolvable questions. A randomized permutation test across seeds and datasets shows that MCQ-DrGRPO achieves significantly higher process and answer accuracy than the baseline. Table 4 shows that the results also hold for multimodal reasoning and for an out-of-domain setting, where the model is trained on geo-guessing and evaluated on year-guessing.

Table 2: **Process-Accuracy (P-Acc)** and **Answer-Accuracy (A-Acc)** for CoTs sampled by Llama3 1B and RL-tuned derivations. The AQuA dataset is used for RL and we additionally evaluate on the MATH and GSM8k datasets, both adapted to the multiple-choice format. The RL training is repeated three times with different random seeds and mean \pm std is reported. P-Acc and A-Acc are calculated using 200 and 2k questions, respectively. Nominal best values are bold.

	AQuA		MATH		GSM8K	
	P-Acc	A-Acc	P-Acc	A-Acc	P-Acc	A-Acc
Base Model (Llama3 1B)	47.0	41.9	45.7	45.1	66.0	55.6
DrGRPO	63.7 \pm 1.6	58.9 \pm 0.0	46.3 \pm 1.5	62.5 \pm 0.1	71.2 \pm 0.8	77.4 \pm 0.1
MCQ-DrGRPO (ours)	65.0 \pm 0.7	59.5 \pm 0.3	50.0 \pm 1.8	62.6 \pm 0.1	73.5 \pm 1.4	78.2 \pm 0.3

Table 3: **Process-Accuracy (P-Acc)** and **Answer-Accuracy (A-Acc)** for CoTs sampled by Qwen2.5 1.5B and RL-tuned derivations. The AQuA dataset is used for RL and we additionally evaluate on the MATH and GSM8k datasets, both adapted to the multiple-choice format. The RL training is repeated three times with different random seeds and mean \pm std is reported. P-Acc and A-Acc are calculated using 200 and 2k questions, respectively. Nominal best values are bold.

	AQuA		MATH		GSM8K	
	P-Acc	A-Acc	P-Acc	A-Acc	P-Acc	A-Acc
Base Model (Qwen2.5 1.5B)	70.0	52.0	68.5	64.7	80.0	66.3
DrGRPO	81.3 \pm 2.1	70.6 \pm 0.1	76.0 \pm 2.2	71.0 \pm 0.2	86.2 \pm 1.8	77.0 \pm 0.2
MCQ-DrGRPO (ours)	82.0 \pm 1.5	71.1 \pm 0.2	76.3 \pm 1.3	72.1 \pm 0.3	87.8 \pm 3.2	81.4 \pm 0.4

5.5 ANALYSIS

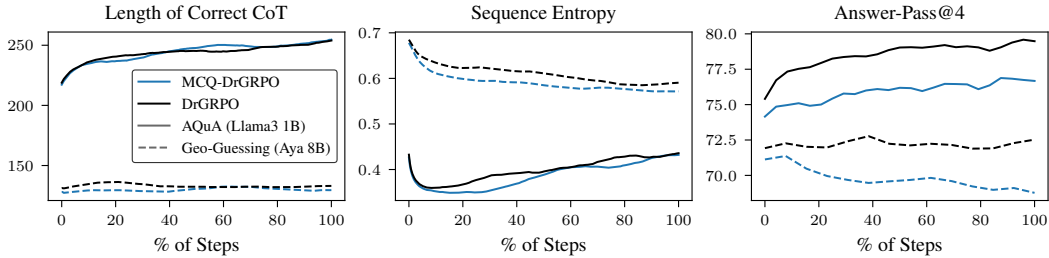


Figure 5: Additional metrics recorded during the reinforcement learning experiments. **Length of Correct CoT**: Average number of tokens in an answer-correct CoT. **Sequence Entropy**: Summed token entropy for a CoT sequence, normalized by length. **Answer-Pass@4**: Percentage of questions with at least one answer-correct CoT among four samples.

In Figure 6, we vary the number of CoTs sampled per question during training. Each data point represents the average of nine values: three models trained with different random seeds, each evaluated on the three math reasoning datasets. The improvement in answer accuracy is consistent across number of samples and the gap in process accuracy is widening with more samples. This can be explained by the clearer identification of unsolvable questions with an increasing number of samples G , as shown in Figure 1 (second graph). To better understand MCQ-DrGRPO’s impact, we track additional metrics during training. The plot on the left of Figure 5 shows that the average length of answer-correct CoTs is comparable for both models, ruling out length bias as an expla-

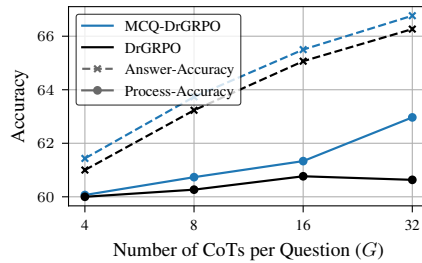


Figure 6: Results for varying number of CoTs per question. Average across three datasets and three random seeds. Relates to Table 2 and Figure 4 (left).

Table 4: **Process-Accuracy (P-Acc)** and **Answer-Accuracy (A-Acc)** for CoTs sampled by the base model (Aya 8B) and RL-tuned derivations. Training on geo-guessing and testing on year-guessing (and vis versa) constitutes out-of-domain evaluation. The RL training is repeated three times with different random seeds and mean \pm std is reported. P-Acc and A-Acc are calculated using 200 and 3.5k questions, respectively. Nominal best values are bold.

	RL Dataset	Geo-Guessing		Year-Guessing	
		P-Acc	A-Acc	P-Acc	A-Acc
Base Model		47.0	46.4	48.5	41.9
DrGRPO	Geo-Guessing	50.2 \pm 2.5	55.6 \pm 0.1	56.8 \pm 3.9	42.6 \pm 0.2
MCQ-DrGRPO (ours)		55.2 \pm 0.6	56.0 \pm 0.3	58.3 \pm 3.1	42.6 \pm 0.1
DrGRPO	Year-Guessing	57.7 \pm 4.2	46.0 \pm 0.2	52.7 \pm 1.0	52.8 \pm 0.3
MCQ-DrGRPO (ours)		59.0 \pm 0.7	46.9 \pm 0.3	57.5 \pm 2.9	53.4 \pm 0.4

nation for the difference in process accuracy. The plot in the middle reveals that the sequence entropy is lower for CoTs generated by the MCQ-DrGRPO model. This means that the output distribution learned by MCQ-DrGRPO is sharper than that learned by DrGRPO. This observation is supported by the graph on the right that shows the percentage of questions that are answered correctly by at least one out of four sampled CoTs. DrGRPO outperforms MCQ-DrGRPO in this metric, which means that the variance of answers is higher using DrGRPO. This aligns with recent work which finds that RL with verifiable rewards does not truly learn new things, but sharpens the distribution toward answer-correct CoTs (Yue et al., 2025). MCQ-DrGRPO effectively prioritizes reliable training signal over diverse but potentially noisy signal. This trade-off results in a sharpened distribution that generates correct CoTs more consistently.

6 RELATED WORK

CoT Process-Correctness The evaluation of CoT reasoning is primarily focused on answer correctness (Wei et al., 2022; Wang et al., 2023b; Fu et al., 2023; Liu et al., 2023; DeepSeek-AI et al., 2025). Because a correct answer does not imply correct reasoning (Wang et al., 2023a), evaluating the process is of interest for many applications (Singhal et al., 2022; Blair-Stanek et al., 2023; Macina et al., 2023). Process correctness in compositional reasoning tasks can often be verified by a parser (Cobbe et al., 2021; Willig et al., 2022; Lyu et al., 2023; Xu et al., 2024), but most natural language tasks require human annotators (Collins et al., 2022; Zelikman et al., 2022; Uesato et al., 2023; Mondorf & Plank, 2024). Only recently have studies explored training models (Golovneva et al., 2023; Prasad et al., 2023) or using LLMs (He et al., 2024; Hao et al., 2024; Bavaresco et al., 2025) to judge the correctness of reasoning chains. Uesato et al. (2023) show that process reward models (PRM), learned from human annotations, are improving process correctness at test time. Recently, PRM training moved away from human annotations (Lightman et al., 2024) towards implicit step-level feedback derived from final answer correctness (Yuan et al., 2025; Wang et al., 2024).

Advantage Calculation and Data Difficulty There are many works that modify the advantage calculation of GRPO (Shao et al., 2024). DrGRPO (Liu et al., 2025) drops the normalization by standard deviation in order to reduce the "question-level difficulty bias". Other works incorporate an entropy reward to encourage more diverse CoTs (Zhang et al., 2025b; Cheng et al., 2025) or penalize uncertainty (Chen et al., 2025). Zhang & Zuo (2025) reweigh the advantage based on question difficulty, calculated as the fraction of correct answers. They increase the weight of CoTs that correctly answer a difficult question. This contrasts the trade-off between difficulty and novelty (Swayamdipta et al., 2020).

7 CONCLUSION

We explicitly modeled the ability of an LLM with CoT reasoning to solve a certain multiple-choice question. To this end, a group of sampled CoTs is used to estimate the probability that the true

performance of the LLM exceeds random guessing. We incorporated the estimated solvability of a question into the objective of an outcome-based reward model and reinforcement learning with group-relative advantage estimation. Experiments on different base models and datasets showed improved process accuracy of emitted CoTs, and additionally improved answer accuracy in the case of RL. Supporting experiments confirmed that answer-correct CoTs from groups with few correct answers are more likely to be process-incorrect and provide noisy learning signal. The method does not introduce additional hyperparameter and requires negligible computation overhead. Taking solvability into account should become the default when using multiple-choice questions for reasoning training. Although it is not straightforward to adapt our solvability formulation to open-answer tasks, one possible direction is reformulation to MCQA (Zhang et al., 2025a). Our work highlights that unsolvable questions are a source of *false positive* CoTs and a method to mitigate this phenomenon in the MCQA setting.

REPRODUCIBILITY

The supplementary material includes training code, data and outputs of our experiments. We will additionally release model checkpoints in a github repository. The Appendix lists model hyperparameters and used prompts.

REFERENCES

- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. URL <https://aclanthology.org/2025.acl-short.20/>.
- Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law, ICAIL ’23*, pp. 22–31, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701979. doi: 10.1145/3594536.3595163. URL <https://doi.org/10.1145/3594536.3595163>.
- Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy enhanced grpo for uncertainty-aware policy optimization, 2025. URL <https://arxiv.org/abs/2505.12346>.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective on reinforcement learning for llms, 2025. URL <https://arxiv.org/abs/2506.14758>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukas Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Katherine M. Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B. Tenenbaum. Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks, 2022. URL <https://arxiv.org/abs/2205.05718>.
- Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün,

- and Sara Hooker. Aya vision: Advancing the frontier of multilingual multimodality, 2025. URL <https://arxiv.org/abs/2505.08751>.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjin Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-rl: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
- Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://openreview.net/forum?id=Fkckkr3ya8>.
- Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A continuous effort to measure large language models’ reasoning performance, 2023. URL <https://arxiv.org/abs/2305.17306>.
- Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=xYlJRpzZtsY>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra,

Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Conguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan

- Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma, Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhiting Hu. LLM reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=b0y6fbSUG0>.
- Hangfeng He, Hongming Zhang, and Dan Roth. SocREval: Large language models with the socratic method for reference-free reasoning evaluation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2736–2764, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.175. URL <https://aclanthology.org/2024.findings-naacl.175/>.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=7Bywt2mQsCe>.
- Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandaraao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Second Conference on Language Modeling (COLM)*, 2025. URL <https://openreview.net/forum?id=90UrTTxp50>.
- Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. *arXiv*, abs/2311.14648, 2024. URL <https://arxiv.org/abs/2311.14648>.
- Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate. *arXiv*, abs/2509.04664, 2025. URL <https://arxiv.org/abs/2509.04664>.
- Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free! In *ICLR 2019 Deep Reinforcement Learning meets Structured Prediction Workshop*, 2019.
- Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth J.F. Jones. The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24(1):93–96, 2017. doi: 10.1109/MMUL.2017.9.
- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. NuminaMath. [<https://huggingface>.

- co/AI-MO/NuminaMath-1.5] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=v8L0pN6EOi>.
- Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*, 2017.
- Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4, 2023. URL <https://arxiv.org/abs/2304.03439>.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding rl-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *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)*, pp. 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.20. URL <https://aclanthology.org/2023.ijcnlp-main.20/>.
- Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5602–5621, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.372. URL <https://aclanthology.org/2023.findings-emnlp.372/>.
- Philipp Mondorf and Barbara Plank. Comparing inferential strategies of humans and large language models in deductive reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9370–9402, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.508. URL <https://aclanthology.org/2024.acl-long.508/>.
- Eric Müller, Matthias Springstein, and Ralph Ewerth. “when was this picture taken?” – image date estimation in the wild. In Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngöve, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (eds.), *Advances in Information Retrieval*, pp. 619–625, Cham, 2017. Springer International Publishing. ISBN 978-3-319-56608-5.
- Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv, abs/2112.00114*, 2021. URL <https://arxiv.org/abs/2112.00114>.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux,

Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameesh Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.

OpenStreetMap Foundation. Nominatim: Open source geocoding with OpenStreetMap data, 2009. URL <https://nominatim.org/>.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL <https://proceedings.mlr.press/v174/pal22a.html>.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.882. URL <https://aclanthology.org/2024.findings-emnlp.882/>.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reasoning chains via correctness and informativeness. In Houda Bouamor, Juan Pino, and Kalika Bali

- (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10066–10086, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.622. URL <https://aclanthology.org/2023.emnlp-main.622/>.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL <http://arxiv.org/abs/1707.06347>.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge, 2022. URL <https://arxiv.org/abs/2212.13138>.
- Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746. URL <https://aclanthology.org/2020.emnlp-main.746/>.
- Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016. ISSN 0001-0782. doi: 10.1145/2812802. URL <https://doi.org/10.1145/2812802>.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.

- Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Yamamoto Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-based and outcome-based feedback, 2023. URL <https://openreview.net/forum?id=MND1kmmNy0O>.
- Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–2739, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.153. URL <https://aclanthology.org/2023.acl-long.153/>.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhi-fang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL <https://aclanthology.org/2024.acl-long.510/>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Chain-of-probe: Examining the necessity and accuracy of CoT step-by-step. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2586–2606, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.140. URL <https://aclanthology.org/2025.findings-naacl.140/>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.
- Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang Wang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms, 2025. URL <https://arxiv.org/abs/2506.14245>.
- Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality?, 2022. URL <https://arxiv.org/abs/2206.10591>.
- Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13326–13365, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.720. URL <https://aclanthology.org/2024.acl-long.720/>.
- Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=8ThnPFhGm8>.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *2nd AI for Math Workshop @ ICML 2025*, 2025. URL <https://openreview.net/forum?id=upehLVgqlb>.

- Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_3ELRdg2sgI.
- Jixiao Zhang and Chunsheng Zuo. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. *arXiv preprint arXiv:2504.09696*, 2025.
- Mengyu Zhang, Xubo Liu, Siyu Ding, Weichong Yin, Yu Sun, Hua Wu, Wenya Guo, and Ying Zhang. Auditable-choice reframing unlocks rl-based verification for open-ended tasks, 2025a. URL <https://arxiv.org/abs/2511.02463>.
- Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity, 2025b. URL <https://arxiv.org/abs/2507.21848>.
- Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. Multiple-choice questions are efficient and robust llm evaluators, 2024.

APPENDIX

A GEO-GUESSING AND YEAR-GUESSING MCQA

We introduce two multiple-choice question answering (MCQA) datasets for multimodal reasoning. Both are derived from the YFCC100M dataset (Thomee et al., 2016) that provides 100M images from Flickr, partially annotated with metadata like tags, title and geolocation. In the following, we describe the construction of the Geo-Guessing MCQA and Year-Guessing MCQA datasets. For Geo-Guessing MCQA, we start with MP16 (Larson et al., 2017), a YFCC100M subset of 7M images that are tagged with geographic coordinates in the form of latitude and longitude. We use Nominatim (OpenStreetMap Foundation, 2009) to translate the coordinates into a textual description of the region where an image was taken. We downloaded each image in the highest available resolution, sorted them by resolution and selected the top 100k instances. These instances are then split into 92k training, 4k development and 4k test instances. The process for Year-Guessing MCQA is the same as for geo, except that we start with DEW (Müller et al., 2017), also a subset of YFCC100M, and extract the year an image was taken from DEW’s additional annotation. They derived the year an image was taken from user-provided metadata like title, description and tags. An example instance of Geo-Guessing MCQA is shown in Figure 7 and of Year-Guessing MCQA in Figure 8.

A.1 GENERATING ANSWER CHOICES

Generating incorrect answer choices (distractors) for MCQA should be handled with care to avoid exposing the correct answer via subtle bias. For example, generating distractor years via symmetric error, e.g., $uniform(year - distance, year + distance)$, let’s a model learn to predict the median value as correct. On the other hand, to have strong distractors, we need to generate choices that are close to the correct answer. To this end, we designed an algorithm that is not biased by the correct value and offers parameters to regulate the maximum distance to the correct value as well as the minimum distance between answers. The latter is needed to increase distinguishability. The algorithm is laid out in Algorithm 1 and the parameters for Year-Guessing are set to $n = 2, d = 20, s = 4$. The algorithm first shifts the center in a random direction by a maximum distance of $d/2$. Then it samples the distractors uniformly around the shifted center, with a maximum distance of $d/2$. The process is repeated until n distractors are generated that all satisfy the minimum intra-distance constraint. For Geo-Guessing we modify the algorithm slightly. The correct value is a pair of latitude and longitude, the value is shifted by a distance in kilometers and random bearing ($uniform(1, 360)$). Distractors that have the same textual representation (by Nominatim) as any of the existing answers are considered equal, even if they have different coordinates. The parameters for Geo-Guessing are $n = 3, d = 2000, s = 5$.

Algorithm 1 Generate Distractor Values with Distance Constraints

Require: correct value a , number of distractors n , maximum distance to correct value d , minimum distance between answers s

- 1: $shifted \leftarrow \text{RANDOMINT}(a - d/2, a + d/2)$
- 2: $answers \leftarrow \{a\}$
- 3: **while** $|answers| < n + 1$ **do**
- 4: $distractor \leftarrow \text{RANDOMINT}(shifted - d/2, shifted + d/2)$
- 5: $valid \leftarrow \text{true}$
- 6: **for each** $value \in answers$ **do**
- 7: **if** $|value - distractor| < s$ **then**
- 8: $valid \leftarrow \text{false}$
- 9: **end if**
- 10: **end for**
- 11: **if** $valid$ **then**
- 12: $answers \leftarrow answers \cup \{distractor\}$
- 13: **end if**
- 14: **end while**
- 15: **output** $answers$



Question: Can you deduce the country, region and city shown in this photograph?

Options: A. Amsterdam, North Holland, Netherlands B. Scotland, United Kingdom
C. High Peak, Derbyshire, East Midlands, England, United Kingdom D. Risør, Agder, Norway

Gold Answer: A

DrGRPO Chain-of-Thought:

This image shows a bustling train station with multiple tracks, cranes, and urban buildings around it. The architecture and style of the buildings, along with the presence of trains and trams, suggest a European city with a well-developed public transport system. The architecture strongly hints at a city in the Netherlands, particularly Amsterdam, known for its canal system and historic buildings. The presence of cranes and the layout of the train station are consistent with the infrastructure found in major Dutch cities. **Additionally, the signage in the image includes the name "Munsters," which is a local landmark in Amsterdam.** The surrounding urban landscape and the style of the buildings further support this deduction.

Judge Label: Incorrect

MCQ-DrGRPO Chain-of-Thought:

This image shows a busy railway station with multiple tracks, a large archway, and urban development in the background. The presence of trains, cranes, and a variety of buildings suggests a major transportation hub in a densely populated area. The architecture and infrastructure styles are consistent with those found in the Netherlands, particularly in Amsterdam. The station's layout and the surrounding buildings resemble those found in Amsterdam's central train station and its surrounding areas. Together, these elements strongly indicate that the photograph was taken in Amsterdam, in the province of North Holland, within the Netherlands.

Judge Label: Correct

Figure 7: An instance of the Geo-Guessing MCQA dataset with two answer-correct CoTs generated by DrGRPO and MCQ-DrGRPO, respectively. The CoTs are evaluated for process-correctness by an LLM judge. The text marked in red is a hallucination flagged by the judge.



Question: From what you can observe, during which period was this image likely captured?

Options: A. 1953 B. 1961 C. 1945

Gold Answer: B

DrGRPO Chain-of-Thought:

This image appears to depict a rugby match, given the attire of the players and the equipment used. The style of the uniforms and the overall aesthetic of the scene suggest it might be from a post-World War II era, as the sport saw a resurgence in popularity **during the 1940s and 1950s. The image seems to capture the spirit of the game during this time**, with players engaged in the sport and a crowd gathered to watch. The presence of spectators in uniforms and the layout of the field are indicative of organized, amateur, or possibly collegiate-level matches that were common during this period.

Judge Label: Incorrect

MCQ-DrGRPO Chain-of-Thought:

The image captures a rugby match with players wearing traditional uniforms and a clear emphasis on teamwork and physicality. Considering the fashion and the context of the sport, it seems that this image was likely captured in the mid-20th century. The attire of the players, particularly the distinctive black and white striped jerseys and the long trousers, aligns with the style from the 1950s and 1960s. Additionally, the quality and composition of the image suggest it was taken during a time when photography technology was advancing but not yet as sophisticated as today's high-definition cameras. Based on these observations, the most likely period for this image is around 1961.

Judge Label: Correct

Figure 8: An instance of the Year-Guessing MCQA dataset with two answer-correct CoTs generated by DrGRPO and MCQ-DrGRPO, respectively. Both models are trained with the Geo-Guessing dataset. The CoTs are evaluated for process-correctness by an LLM judge. The text marked in red is a hallucination flagged by the judge.

B STATISTICAL SIGNIFICANCE

We perform a stratified random permutation test to establish the significance of the reported results. For each group of experiments, we pool the accuracy values of both methods across all seeds and randomly reassign them to two groups. We repeat this permutation process 100k times within each dataset independently, then compute the mean difference across datasets for each permutation. The two-tailed p-value is the proportion of permutations where the absolute value of the permuted mean difference is greater than or equal to the absolute value of the observed mean difference. The following p-values indicate the statistical significance of the improvement when using the proposed MCQ variants compared to the baseline ORM and RL methods. Table 1 Llama3-1B P-Acc p-value: 0.0115; Table 1 Llama3-8B P-Acc p-value: 0.0010; Table 2 Llama3-1B 32 Samples P-Acc p-value: 0.0077, A-Acc p-value: 0.0008; Table 4 Aya-8B P-Acc p-value: 0.0249, A-Acc p-value: 0.0023.

Table 5: Hyperparameter for reward model training. Hyperparameter were selected by cross-entropy loss on the development set.

	Llama3 1B		Llama3 8B	
	ORM	MCQ-ORM	ORM	MCQ-ORM
Batch Size			512	
Dropout			0.0	
Gradient Norm			1.0	
Learning Rate			0.0001	
Weight Decay			0.001	
Optimizer			AdamW	
Hidden Layer Dimensions	128:4	64:8	128	64
LR Schedule	cosine	linear	cosine	cosine
LR Warmup	0.1	0.05	0.1	0.1

Table 6: Hyperparameter used for reinforcement learning experiments.

Hyperparameter	Llama3 1B	Aya 8B
Train Batch Size		128
Optimizer		AdamW
Max. Gradient Norm		1.0
Learning Rate		0.000005
LR Schedule		constant
Weight Decay		0.0
KL Weight		0.01
Rollout Batch Size		128
Rollouts per Step		128
Rollout Temperature		1.0
Rollout Min. Tokens		64
Rollout Max. Tokens		1024
Eval Temperature		1.0
Eval Max. Tokens		1280
Samples per Rollout	32	8
Max. Prompt Length	1024	2560
Frozen Layers	None	Image Encoder and lower half of LLM Layers

C METADATA FOR MODEL TRAINING

We list the hyperparameter used for reward model training in Table 5. We ran a minimal grid search to find the best hidden dimensions, learning rate, schedule and warmup for the baseline ORM and MCQ-ORM. The hyperparameters were chosen based on the lowest loss on the development set. The hyperparameter for reinforcement learning experiments are listed in Table 6. They were chosen based on initial experiments with DrGRPO and Llama3 1B. The maximum prompt length for multimodal Aya is higher because it includes image tokens. Due to resource constraints, we set the number of samples to 8 and did not update the image encoder as well as the lower half (16) of the LLM layers.

D META EVALUATION

We use GPT-4.1 (OpenAI et al., 2024) with version *gpt-4.1-2025-04-14* as the judge to assess process correctness. To ensure future comparison and reproducibility of our results, we released the verbatim CoTs and the full assessment of the judge. The prompt we used is shown in Figure 9. We further conduct a meta evaluation that compares the LLM judge with human judgments and its ability to detect synthetically corrupted gold CoTs. Golovneva et al. (2023) released a dataset of 200 model generated CoTs for GSM8k together with human judgments of process correctness. We compare these human judgments with those from our LLM judge in Table 7. In 97% of the cases the human and LLM judge agree on the process correctness. Four of the six cases in which the human and judge disagree are shown in Figure 10, Figure 11, Figure 12 and Figure 13. It is up to the reader

Table 7: Human Meta Evaluation GSM8k. 97% overall.

Human	#CoT	Judge Correct	Judge Incorrect	Judge Accuracy
Correct	109	103	6	94.5%
Incorrect	91	0	91	100%

Table 8: Synthetic Meta Evaluation AQuA.

Gold CoT	Corrupted CoT	#Count
Correct	Incorrect	154
Correct	Correct	5
Incorrect	Correct	0
Incorrect	Incorrect	41

to decide whether the concerns by the judge are minor inaccuracies or invalidates the CoT. Next, we evaluate our LLM judge on a synthetic dataset. To this end, we use the gold rationals of AQuA and corrupt them by replacing a randomly selected number with a randomly generated number. Table 8 shows the judgment outcomes for the four possible cases. In the first row, we see the most desired outcome — the gold CoT is judged as correct and the corrupted CoT as incorrect. In only five cases, the LLM judge does not flag the corrupted CoT as incorrect. Because the LLM judge flags some gold rationals as incorrect, we manually inspect those CoTs. We find that all of them contain an error and show an example of an incorrect gold CoT in Figure 14.

Your task is to evaluate whether a reasoning process is fully correct and leads to the ground truth answer. I will present you a question, its ground truth answer and the reasoning process. Please analyze the reasoning and tell me if it is fully correct. Return your final assessment as `\\boxed{{yes}}` or `\\boxed{{no}}`. Remember, the presented answer to the question is the ground truth and you are solely judging the correctness of the reasoning process.

Question:

<>

Correct Answer:

<>

Reasoning:

<>

Is the reasoning process correct? Think step by step and return your final judgement as `\\boxed{{yes}}` or `\\boxed{{no}}`.

Figure 9: Prompt for LLM judge to assess the process correctness of a CoT.

E ORM ANSWER-ACCURACY

In addition to the Process-Accuracy (P-Acc) results in Table 1, we report the Answer-Accuracy (A-Acc) results in Table 9 for the reward model experiments. Here, the ORM baseline outperforms the MCQ-ORM method. However, in practice, one can use the ORM to select the CoT that is most likely answer-correct and subsequently the MCQ-ORM to select one of the answer-correct CoTs that is most likely process-correct.

F LLM USAGE

We used LLMs to rephrase single sentences, but not full paragraphs. We also used them to help with the correct formatting of equations and tables.

Question: The marching band is ordering new uniforms. Each uniform comes with a hat that costs \$25, a jacket that costs three times as much as the hat, and pants that cost the average of the costs of the hat and jacket. How much does each uniform cost total?

Gold Answer: 150

Chain-of-Thought:
 Step 1 - The jacket costs $\$25 * 3 = \75 .
 Step 2 - The average cost of the hat and jacket is $\$25 + \$75 = \$100$.
 Step 3 - The pants cost $\$100 / 2 = \50 .
 Step 4 - So the total cost of each uniform is $\$25 + \$75 + \$50 = \150 .
 Step 5 - A: 150

Human Label: Correct

Judge Label: Incorrect
Judge Reasoning: Let's analyze the reasoning step by step:
 Step 1: The jacket costs $\$25 * 3 = \75 . - This is correct, as the jacket is three times the cost of the hat.
 Step 2: The average cost of the hat and jacket is $\$25 + \$75 = \$100$. - This is not the average; this is the sum. The average should be $(25 + 75) / 2$.
 Step 3: The pants cost $\$100 / 2 = \50 . - This step is correct in that it divides the sum by 2 to get the average, but the previous step mislabels the sum as the average.
 Step 4: So the total cost of each uniform is $\$25 + \$75 + \$50 = \150 . - This is correct, as it adds the costs of the hat, jacket, and pants.
 Step 5: A: 150 - This is the correct answer.
 However, in Step 2, the reasoning says "the average cost of the hat and jacket is $\$25 + \$75 = \$100$," which is not the average but the sum. The average is calculated in the next step, but the labeling is incorrect. Therefore, the reasoning process contains a misstatement about what the average is, even though the calculation ultimately arrives at the correct answer.
 Final assessment: $\boxed{\text{no}}$

Figure 10: Example of disagreement between human and LLM judgment.

Table 9: **Answer-Accuracy (A-Acc)** using different methods of CoT scoring. The task is to score multiple candidate CoTs. The highest scoring CoT is then evaluated for answer-correctness. This is done for all development set questions of the three multiple-choice QA datasets. Outcome-supervised reward models (ORM) are trained with three different random seeds and mean \pm std is reported. Nominal best values are bold.

	Llama3 1B			Llama3 8B		
	AQuA	MATH	GSM8K	AQuA	MATH	GSM8K
Random	41.9	45.1	55.6	77.2	71.7	92.4
Majority Voting	53.1	56.4	73.2	88.5	81.4	98.1
Answer Confidence	59.9	59.8	82.7	88.9	83.3	98.3
ORM	66.3 ± 0.1	63.8 ± 0.3	86.8 ± 0.4	89.4 ± 0.1	83.7 ± 0.1	98.2 ± 0.1
MCQ-ORM (ours)	63.1 ± 0.1	62.7 ± 0.3	85.2 ± 0.2	89.4 ± 0.1	83.6 ± 0.3	98.3 ± 0.0

Question: Katy makes coffee using teaspoons of sugar and cups of water in the ratio of 7:13. If she used a total of 120 teaspoons of sugar and cups of water, calculate the number of teaspoonfuls of sugar she used.

Gold Answer: 42

Chain-of-Thought:

Step 1 - The total ratio representing the number of teaspoons of sugar and cups of water used is $7+13 = 20$.

Step 2 - The fraction representing the number of teaspoons of sugar used is $7/20$, and since she used 120 teaspoons of sugar, she used $7/20 * 120 = 42$ teaspoons of sugar

Step 3 - A: 42

Human Label: Correct

Judge Label: Incorrect

Judge Reasoning: Let's analyze the reasoning step by step:

Step 1: The total ratio representing the number of teaspoons of sugar and cups of water used is $7+13 = 20$. This is correct. The total parts in the ratio are 7 (sugar) + 13 (water) = 20 parts.

Step 2: The fraction representing the number of teaspoons of sugar used is $7/20$, and since she used 120 teaspoons of sugar, she used $7/20 * 120 = 42$ teaspoons of sugar. Here is the issue: The statement says "since she used 120 teaspoons of sugar," but the problem says "she used a total of 120 teaspoons of sugar and cups of water." That is, 120 is the total of both ingredients, not just sugar. The reasoning should be: Number of teaspoons of sugar = $(7/20) \times 120 = 42$.

The calculation is correct, but the reasoning misstates what the 120 represents. It should say "since she used a total of 120 teaspoons of sugar and cups of water," not "since she used 120 teaspoons of sugar."

Step 3: A: 42 The answer is correct.

However, because the reasoning misstates what the 120 represents, the reasoning process is not fully correct.

Final assessment: $\boxed{\text{no}}$

Figure 11: Example of disagreement between human and LLM judgment.

Question: Lee used to be able to run the 400-meter hurdles two seconds faster than Gerald would run the 400-meter hurdles. But Gerald changed his diet, which improved his speed by 10%. If Lee runs the 400-meter hurdles in 38 seconds, how fast can Gerald, with his improved diet, run the 400-meter hurdles, in seconds?

Gold Answer: 36

Chain-of-Thought:

Step 1 - Since Lee is able to run the 400-meter hurdles two seconds faster than Gerald, then Gerald runs the 400-meter hurdles in $38 + 2 = 40$ seconds.

Step 2 - If Gerald's speed improves by 10%, then his time to run the 400-meter hurdles is reduced by $10/100 * 40 = 4$ seconds.

Step 3 - So, Gerald's time to run the 400-meter hurdles, after his diet improvements, is $40 - 4 = 36$ seconds.

Step 4 - A: 36

Human Label: Correct

Judge Label: Incorrect

Judge Reasoning: Let's analyze the reasoning step by step:

Step 1: Lee runs the 400m hurdles in 38 seconds, which is 2 seconds faster than Gerald's old time. So Gerald's old time is $38 + 2 = 40$ seconds. This is correct.

Step 2: Gerald's speed improves by 10%. The reasoning says that his time is reduced by 10% of 40 seconds, i.e., 4 seconds. This is a critical point.

- If speed increases by 10%, then the time to cover the same distance decreases, but not by 10%. The relationship between speed and time is inverse: $\text{time} = \text{distance} / \text{speed}$. - If original speed is S , and new speed is $1.1S$, then new time is $T' = D / (1.1S) = (1/1.1) * (D/S) = (1/1.1) * T \approx 0.909 * T$. - So, the new time should be $40 * (1/1.1) \approx 36.36$ seconds, not 36 seconds.

Step 3: The reasoning subtracts 4 seconds (10% of 40) from 40, getting 36 seconds. This is incorrect, as explained above.

Final assessment: The reasoning process is not fully correct. $\boxed{\text{no}}$

Figure 12: Example of disagreement between human and LLM judgment.

Question: Candice put 80 post-it notes in her purse before she headed out to her job at the coffee shop. On her way, she stopped off at the store and purchased a package of Post-it notes; At work, she placed a single Post-it note on each of 220 different cups of coffee. If she had 23 post-it notes remaining overall, how many Post-it notes were in the package that she purchased?

Gold Answer: 163

Chain-of-Thought:

Step 1 - If she placed one post-it note on each of the 220 coffee cups, she used $220/1 = 220$ post-it notes.

Step 2 - If she had 80 post-it notes in her purse before she purchased the package, then she used $220 - 80 = 140$ post-it notes from the package.

Step 3 - If she had 23 post-it notes remaining, then the package contained $140 + 23 = 163$ post-it notes.

Step 4 - A: 163

Human Label: Correct

Judge Label: Incorrect

Judge Reasoning: Let's analyze the reasoning step by step:

****Step 1:**** "If she placed one post-it note on each of the 220 coffee cups, she used $220/1 = 220$ post-it notes." - This is correct. She used 220 post-it notes.

****Step 2:**** "If she had 80 post-it notes in her purse before she purchased the package, then she used $220 - 80 = 140$ post-it notes from the package." - This is incorrect. She had 80 post-it notes, but after using 220, she would have a deficit of 140 post-it notes, not that she used 140 from the package. She could not have used $220 - 80 = 140$ from the package, because she also had 23 left at the end.

****Step 3:**** "If she had 23 post-it notes remaining, then the package contained $140 + 23 = 163$ post-it notes." - This is also incorrect, because the calculation in Step 2 was incorrect. The correct way is to set up the equation: $80 + x$ (package) - 220 (used) = 23 (remaining), so $80 + x = 220 + 23$, so $x = 163$.

****Step 4:**** "A: 163" - The answer is correct, but the reasoning to get there is not.

****Conclusion:**** The reasoning process is not fully correct, even though the final answer is correct.

Final assessment: `\\boxed{{no}}`

Figure 13: Example of disagreement between human and LLM judgment.

Question: A trader bought a car at 25% discount on its original price. He sold it at a 40% increase on the price he bought it. What percent of profit did he make on the original price?

Options: A. 118 B. 110 C. 105 D. 113 E. 115

Gold Answer: C

Gold CoT: Original price = 100 CP = 75 S = $75 * (140/100) = 105$ **100 - 105 = 5%** Answer: C

Judge Label: Incorrect

Judge Reasoning: Let's analyze the reasoning step by step:

1. Original price = 100

This is a good assumption for calculation.

2. CP = 75

Since the trader bought at a 25% discount, the cost price (CP) is 75.

3. S = $75 * (140/100) = 105$

He sold it at a 40% increase on the price he bought it, so selling price (S) = $75 \times 1.4 = 105$.

4. $100 - 105 = 5\%$

This step is incorrect. The calculation " $100 - 105 = 5\%$ " is not the correct way to find the percent profit on the original price.

The correct way:

Profit = Selling Price - Original Price = $105 - 100 = 5$ Profit percent on original price = $(\text{Profit} / \text{Original Price}) \times 100 = (5 / 100) \times 100 = 5\%$

However, the answer is 5%, which matches option C (105%), but the reasoning is not clearly or correctly explained, especially in the last step.

Conclusion: The reasoning process contains a calculation error in the last step and does not clearly explain how the percent profit is calculated on the original price. Therefore, the reasoning is not fully correct.

`\\boxed{{no}}`

Figure 14: Gold rational of AQuA with error marked in red.