

000 001 002 003 004 005 BOOSTING PROCESS-CORRECT CoT REASONING BY 006 MODELING SOLVABILITY OF MULTIPLE-CHOICE QA 007 008 009

010 **Anonymous authors**
011

012 Paper under double-blind review
013
014
015
016
017
018
019
020
021
022
023
024

ABSTRACT

011 Reasoning quality in large language models depends not only on producing correct
012 answers but also on generating valid intermediate steps. We study this through
013 multiple-choice question answering (MCQA), which provides a controlled setting
014 with fixed answer options. Our analysis shows that when questions are effectively
015 unsolvable for a model, spurious chains of thought (CoTs) are more likely to appear,
016 leading to false positives. By estimating the solvability of each question,
017 we uncover an intermediate regime where learning is most effective. Building
018 on this insight, we adapt outcome-supervised reward models and reinforcement
019 learning with group-relative advantage to incorporate solvability into their objectives.
020 Across experiments on math and multimodal datasets, these modifications
021 consistently yield higher rates of process-correct reasoning and, in reinforcement
022 learning, improved answer accuracy as well. Our results highlight solvability as a
023 key factor for reducing hallucinations and increasing reliability in CoT reasoning.
024

025 1 INTRODUCTION 026

027 In many applications of CoT reasoning, the generated thought process is as important as the final
028 answer. While some tasks provide gold-standard reasoning chains that can effectively be used for
029 supervised training (Nye et al., 2021; Dziri et al., 2023; Hochlehnert et al., 2025), most datasets
030 lack such annotations. For these cases, correct reasoning has to be incentivized by rewards on
031 correct final answers (Wen et al., 2025). It is known that CoTs can lead to the correct answer,
032 despite an incorrect explanation. Grattafiori et al. (2024) note that this often occurs for questions
033 where only a small fraction of the generated answers is correct. In this work, we investigate this
034 observation in controlled experiments on multiple datasets. To avoid confounding factors of noisy
035 answer extraction and matching, we focus on multiple-choice question answering. This format
036 is popular for evaluating models and widely used training sets like NuminaMath (LI et al., 2024)
037 contain a large fraction of multiple-choice questions. The fixed number of answer options also
038 allows us to explicitly model the solvability of a question. We find that unsolvable questions promote
039 false positive CoTs. Additionally, in a controlled finetuning experiment we show that there is a
040 sweet spot of questions for which neither a small nor high fraction of CoTs lead to the correct
041 answer. We make use of these findings by modifying the objective function of an outcome-based
042 reward model (ORM) and by adjusting the advantage calculation of group relative reinforcement
043 learning (RL). The proposed modifications lead to more process-correct CoTs and additionally, in
044 the case of RL, to improved answer accuracy. Furthermore, following the argumentation of Kalai
045 & 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.
046

047 2 BACKGROUND 048

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

$$051 \quad o_{ij} \sim \pi_\theta(\cdot | q_i), \quad \text{where } o_{ij} = (t_{ij}, \hat{y}_{ij}). \quad (1)$$

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

054 results. Given the ground-truth answer y_i , the correctness of the generated answer is determined by
 055 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
 056 open-ended answers can be ambiguous, multiple-choice question answering (MCQA) is a popular
 057 format. Each question includes a letter-indexed list of predefined answer choices c_i with exactly one
 058 correct choice, and scoring reduces to exact letter matching.
 059

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

$$063 \quad \text{A-Acc} := \frac{1}{|\mathbb{D}|G} \sum_{i=1}^{|\mathbb{D}|} \sum_{j=1}^G \mathbb{1}[y_i = \hat{y}_{ij}]. \quad (2)$$

066 A CoT consists of the thought process and the final answer. The correctness of the latter can be deter-
 067 mined by comparing it to the ground-truth answer. Because there does not exist *the one* ground-truth
 068 thought process, we employ an LLM to judge its correctness. Although this is common practice,
 069 it has to be handled with care (He et al., 2024; Hao et al., 2024; Bavaresco et al., 2025). As such,
 070 we conduct an extensive meta-evaluation of the judge by reporting correlation with human judg-
 071 ments, measuring performance on a synthetic dataset, and manually evaluating a subset of the judg-
 072 ments (Appendix D). In addition, we release all outputs verbatim to facilitate future comparisons
 073 with our work. Formally, the judge \mathcal{J}_{LLM} receives the question q_i , a thought process t_{ij} , and the
 074 ground-truth answer y_i as input, and returns the binary judgment:
 075

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

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

$$077 \quad \text{P-Acc} := \frac{1}{|\mathbb{D}_{AC}|} \sum_{q_i, t_{ij}, y_i \in \mathbb{D}_{AC}} \mathcal{J}_{LLM}(q_i, t_{ij}, y_i). \quad (4)$$

079 In this work, we report process accuracy solely on the subset of answer-correct CoTs, denoted as
 080 \mathbb{D}_{AC} . A CoT that is answer-correct but process-incorrect is referred to as *false positive*.
 081

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

$$085 \quad \mathcal{L}_{ij}^{\text{BCE}}(\pi_\phi) = -z_{ij} \log \pi_\phi(h_{ij}) - (1 - z_{ij}) \log(1 - \pi_\phi(h_{ij})) \quad (5)$$

087 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
 088 sampled output (q_i, o_{ij}) , e.g., the raw text or the LLM’s last hidden state during generation. The
 089 ORM can then be used to rerank outputs at test-time or as reward estimator in reinforcement learning.

090 **Reinforcement Learning with Group Relative Advantage** Recently, reinforcement learning
 091 with estimation of group relative advantage has gained renewed traction, especially in domains with
 092 verifiable rewards (Kool et al., 2019; Shao et al., 2024; DeepSeek-AI et al., 2025). The approach
 093 eliminates the complexity of training a reward model and the accompanying problems such as re-
 094 ward hacking or data bias. Instead, the advantage or value of an action o_{ij} is determined relative
 095 to other samples for the same input. This means we sample multiple outputs per question and the
 096 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 ex-
 097 tracted from o_{ij} , e.g. due to incorrect answer format. The advantage is then calculated relative to
 098 the rewards of the other samples. Specifically, the formulations of the GRPO (Shao et al., 2024) and
 099 DrGRPO (Liu et al., 2025) variants are:

$$100 \quad A_{ij}^{\text{GRPO}} = \frac{1}{\sigma(r_i)} (r_{ij} - \frac{1}{G} \sum_{k=1}^G r_{ik}) \quad \text{and} \quad A_{ij}^{\text{DrGRPO}} = r_{ij} - \frac{1}{G} \sum_{k=1}^G r_{ik}, \quad (6)$$

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

$$105 \quad \nabla_\theta J(\theta) \approx \frac{1}{G} \sum_{j=1}^G A_{ij} \nabla_\theta \log \pi_\theta(o_{ij} \mid q_i), \quad (7)$$

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

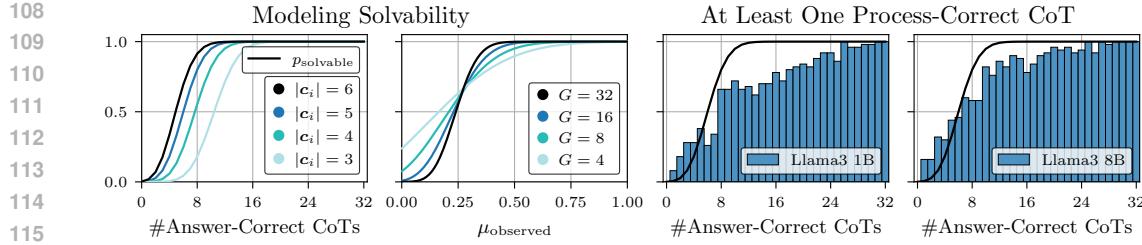


Figure 1: **Modeling Solvability:** The probability that a question is solvable by a given model, as defined by Equation 11. (Left) Varying number of answer options for the multiple-choice question. (Right) Varying number of sampled CoTs per question. **At Least One Process-Correct CoT:** Fraction of 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 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 Beta(α_i, β_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:

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

The left section of Figure 1 illustrates $p_{\text{solv}}^\theta(q_i)$ 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 $p_{\text{solv}}^\theta(q_i)$ 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.

162
 163 Table 1: **Process-Accuracy (P-Acc)** using different methods of CoT scoring. The task is to score
 164 multiple candidate CoTs that all lead to the correct answer. The highest scoring CoT is then eval-
 165 uated for process-correctness. This is done for 200 questions of the three multiple-choice QA
 166 datasets. Oracle gives the upper bound because not every candidate set contains a process-correct
 167 CoT. Outcome-supervised reward models (ORM) are trained with three different random seeds and
 168 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 ± 0.8 | 64.5 ± 1.5 | 87.8 ± 1.0 | 90.3 ± 0.2 | 81.0 ± 0.8 | 95.7 ± 0.2 |
| MCQ-ORM (ours) | 70.0 ± 0.7 | 65.7 ± 0.2 | 88.7 ± 0.5 | 92.0 ± 0.4 | 83.5 ± 0.4 | 96.2 ± 0.5 |

4 BOOSTING PROCESS-CORRECTNESS AT TEST-TIME

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

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

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

4.1 EXPERIMENTS

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

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

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

216 we sample 32 CoTs and determine the predicted answer by majority vote. The subset of CoTs that
 217 lead to the majority answer is the candidate set. The task is to select one CoT among the candidates
 218 that is most likely process-correct. Each method is evaluated on the same 200 questions from each
 219 dataset ranging from 6 to 32 candidate CoTs. Because it is not guaranteed that at least one candidate
 220 is process-correct, we also report oracle results as an upper bound.
 221

222 4.2 RESULTS

223 The results in Table 1 show that both base models are generally capable of generating process-
 224 correct CoTs. The large gap between the random baseline and the oracle shows substantial room to
 225 test the capabilities of the considered selection methods. The CoPS (Wang et al., 2025) and faith-
 226 fulness (Paul et al., 2024) baselines are making use of early answer probing. After each reasoning
 227 step, the model is forced to decode the correct answer letter and its token probability is recorded in
 228 a vector. This early answer probability vector is then used to draw conclusions about the reasoning
 229 process. CoPS estimates the quality of a CoT by considering the average probability of early answers
 230 and their increase over time. The CoT faithfulness metric was developed to measure the alignment
 231 of the model’s internal with its external textual reasoning. It is defined as the area over the curve of
 232 the early answering probability vector. We can see that both of these metrics are outperformed by
 233 a simple baseline that ranks CoTs by final answer confidence, i.e., the probability assigned to the
 234 correct answer letter. Using the score of a reward model to select the best CoT largely outperforms
 235 the aforementioned baselines. Our proposed MCQ-ORM that takes the solvability of a question into
 236 account consistently outperforms the unmodified ORM across considered datasets and base models.
 237 Although the effect size is small, a random permutation test (Appendix B) shows that the results are
 238 overall significant.
 239

240 5 REINFORCEMENT LEARNING WITH ADJUSTED ADVANTAGE

241

242 We start with an analysis of advantage values calculated by GRPO and DrGRPO. The plots on the
 243 left and middle of Figure 2 show that a sample o_{ij} with positive reward $r_{ij} = 1$ gets the highest
 244 individual advantage if all other samples in the group received a negative reward. Comparing
 245 this to the right section of Figure 1, we see that these samples correspond to CoTs that are most
 246 likely process-incorrect. To further investigate the impact of this advantage shape, we conduct a
 247 controlled experiment where we estimate the learning potential of question-CoT pairs, depending
 248 on the number of answer-correct CoTs in the group.
 249

250

5.1 ESTIMATING LEARNING POTENTIAL

251

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

256

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

257

258

259 For each bucket, such a dataset with $2k$ instances is sampled and used to finetune the base model.
 260 We then evaluate the finetuned models on the development set of the respective dataset and
 261 measure the improvement in answer-accuracy over the base model. Figure 3 shows the results for
 262 AQuA (Ling et al., 2017), MedMCQA (Pal et al., 2022) and SocialIQA (Sap et al., 2019) using
 263 Llama2 7B (Touvron et al., 2023). Each experiment is repeated five times with different random
 264 seeds and the mean is depicted. The seed affects the subset selection and the order of training
 265 batches. Buckets with less than $2k$ questions are merged with their neighbor bucket. For all
 266 datasets, the observed distribution of accuracy improvement is left-skewed, with a linear decrease
 267 to the right and a steep increase on the left. The position of the distribution mode varies depending
 268 on the dataset. In the following, we derive a simple model that describes the observed accuracy
 269 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}) .

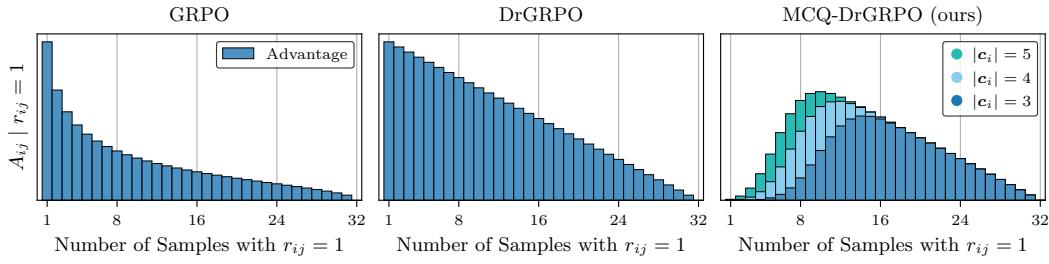


Figure 2: Advantage values of a single CoT with positive reward. 32 CoTs are sampled for each question and the x-axis denotes the number of answer-correct (positive reward) CoTs in a group. MCQ-DrGRPO down-weights CoTs that are generated for unsolvable questions. The probability that a multiple-choice 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.

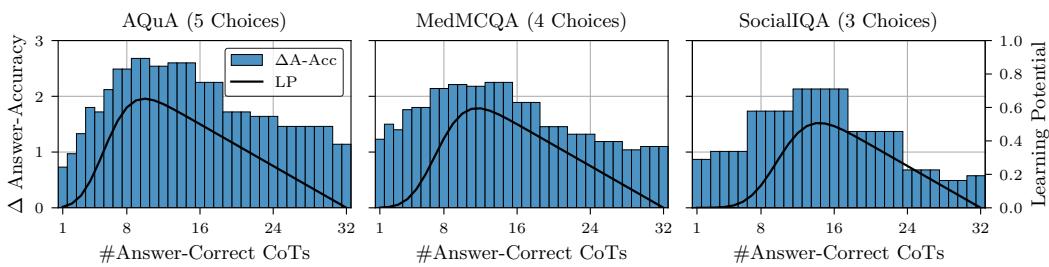


Figure 3: We sample 32 CoTs for each question in the respective training set. Questions are then categorized into buckets based on the number of answer-correct CoTs. We randomly sample questions from each bucket and pair them with exactly one of their answer-correct CoTs. We finetune the base model on 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:

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

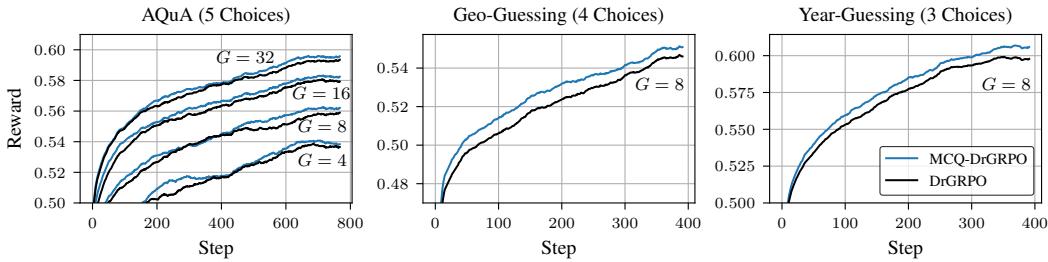


Figure 4: The average reward during RL training with DrGRPO and MCQ-DrGRPO. The math reasoning dataset AQuA is used to train Llama3 1B and geo/year-guessing datasets are used to train multimodal Aya 8B. First graph also shows ablations for different numbers of sampled CoTs per question. Each model is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

378
 379
 380
 381
 382
 383
 384
 385
 386
 387
 388
 389
 390
 391
 392
 393
 394
 395
 396
 397
 398
 399
 400
 401
 402
 403
 404
 405
 406
 407
 408
 409
 410
 411
 412
 413
 414
 415
 416
 417
 418
 419
 420
 421
 422
 423
 424
 425
 426
 427
 428
 429
 430
 431
 432
 433
 434
 435
 436
 437
 438
 439
 440
 441
 442
 443
 444
 445
 446
 447
 448
 449
 450
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460
 461
 462
 463
 464
 465
 466
 467
 468
 469
 470
 471
 472
 473
 474
 475
 476
 477
 478
 479
 480
 481
 482
 483
 484
 485
 486
 487
 488
 489
 490
 491
 492
 493
 494
 495
 496
 497
 498
 499
 500
 501
 502
 503
 504
 505
 506
 507
 508
 509
 510
 511
 512
 513
 514
 515
 516
 517
 518
 519
 520
 521
 522
 523
 524
 525
 526
 527
 528
 529
 530
 531
 532
 533
 534
 535
 536
 537
 538
 539
 540
 541
 542
 543
 544
 545
 546
 547
 548
 549
 550
 551
 552
 553
 554
 555
 556
 557
 558
 559
 560
 561
 562
 563
 564
 565
 566
 567
 568
 569
 570
 571
 572
 573
 574
 575
 576
 577
 578
 579
 580
 581
 582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593
 594
 595
 596
 597
 598
 599
 600
 601
 602
 603
 604
 605
 606
 607
 608
 609
 610
 611
 612
 613
 614
 615
 616
 617
 618
 619
 620
 621
 622
 623
 624
 625
 626
 627
 628
 629
 630
 631
 632
 633
 634
 635
 636
 637
 638
 639
 640
 641
 642
 643
 644
 645
 646
 647
 648
 649
 650
 651
 652
 653
 654
 655
 656
 657
 658
 659
 660
 661
 662
 663
 664
 665
 666
 667
 668
 669
 670
 671
 672
 673
 674
 675
 676
 677
 678
 679
 680
 681
 682
 683
 684
 685
 686
 687
 688
 689
 690
 691
 692
 693
 694
 695
 696
 697
 698
 699
 700
 701
 702
 703
 704
 705
 706
 707
 708
 709
 710
 711
 712
 713
 714
 715
 716
 717
 718
 719
 720
 721
 722
 723
 724
 725
 726
 727
 728
 729
 730
 731
 732
 733
 734
 735
 736
 737
 738
 739
 740
 741
 742
 743
 744
 745
 746
 747
 748
 749
 750
 751
 752
 753
 754
 755
 756
 757
 758
 759
 760
 761
 762
 763
 764
 765
 766
 767
 768
 769
 770
 771
 772
 773
 774
 775
 776
 777
 778
 779
 780
 781
 782
 783
 784
 785
 786
 787
 788
 789
 790
 791
 792
 793
 794
 795
 796
 797
 798
 799
 800
 801
 802
 803
 804
 805
 806
 807
 808
 809
 810
 811
 812
 813
 814
 815
 816
 817
 818
 819
 820
 821
 822
 823
 824
 825
 826
 827
 828
 829
 830
 831
 832
 833
 834
 835
 836
 837
 838
 839
 840
 841
 842
 843
 844
 845
 846
 847
 848
 849
 850
 851
 852
 853
 854
 855
 856
 857
 858
 859
 860
 861
 862
 863
 864
 865
 866
 867
 868
 869
 870
 871
 872
 873
 874
 875
 876
 877
 878
 879
 880
 881
 882
 883
 884
 885
 886
 887
 888
 889
 890
 891
 892
 893
 894
 895
 896
 897
 898
 899
 900
 901
 902
 903
 904
 905
 906
 907
 908
 909
 910
 911
 912
 913
 914
 915
 916
 917
 918
 919
 920
 921
 922
 923
 924
 925
 926
 927
 928
 929
 930
 931
 932
 933
 934
 935
 936
 937
 938
 939
 940
 941
 942
 943
 944
 945
 946
 947
 948
 949
 950
 951
 952
 953
 954
 955
 956
 957
 958
 959
 960
 961
 962
 963
 964
 965
 966
 967
 968
 969
 970
 971
 972
 973
 974
 975
 976
 977
 978
 979
 980
 981
 982
 983
 984
 985
 986
 987
 988
 989
 990
 991
 992
 993
 994
 995
 996
 997
 998
 999
 1000
 1001
 1002
 1003
 1004
 1005
 1006
 1007
 1008
 1009
 1010
 1011
 1012
 1013
 1014
 1015
 1016
 1017
 1018
 1019
 1020
 1021
 1022
 1023
 1024
 1025
 1026
 1027
 1028
 1029
 1030
 1031
 1032
 1033
 1034
 1035
 1036
 1037
 1038
 1039
 1040
 1041
 1042
 1043
 1044
 1045
 1046
 1047
 1048
 1049
 1050
 1051
 1052
 1053
 1054
 1055
 1056
 1057
 1058
 1059
 1060
 1061
 1062
 1063
 1064
 1065
 1066
 1067
 1068
 1069
 1070
 1071
 1072
 1073
 1074
 1075
 1076
 1077
 1078
 1079
 1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092
 1093
 1094
 1095
 1096
 1097
 1098
 1099
 1100
 1101
 1102
 1103
 1104
 1105
 1106
 1107
 1108
 1109
 1110
 1111
 1112
 1113
 1114
 1115
 1116
 1117
 1118
 1119
 1120
 1121
 1122
 1123
 1124
 1125
 1126
 1127
 1128
 1129
 1130
 1131
 1132
 1133
 1134
 1135
 1136
 1137
 1138
 1139
 1140
 1141
 1142
 1143
 1144
 1145
 1146
 1147
 1148
 1149
 1150
 1151
 1152
 1153
 1154
 1155
 1156
 1157
 1158
 1159
 1160
 1161
 1162
 1163
 1164
 1165
 1166
 1167
 1168
 1169
 1170
 1171
 1172
 1173
 1174
 1175
 1176
 1177
 1178
 1179
 1180
 1181
 1182
 1183
 1184
 1185
 1186
 1187
 1188
 1189
 1190
 1191
 1192
 1193
 1194
 1195
 1196
 1197
 1198
 1199
 1200
 1201
 1202
 1203
 1204
 1205
 1206
 1207
 1208
 1209
 1210
 1211
 1212
 1213
 1214
 1215
 1216
 1217
 1218
 1219
 1220
 1221
 1222
 1223
 1224
 1225
 1226
 1227
 1228
 1229
 1230
 1231
 1232
 1233
 1234
 1235
 1236
 1237
 1238
 1239
 1240
 1241
 1242
 1243
 1244
 1245
 1246
 1247
 1248
 1249
 1250
 1251
 1252
 1253
 1254
 1255
 1256
 1257
 1258
 1259
 1260
 1261
 1262
 1263
 1264
 1265
 1266
 1267
 1268
 1269
 1270
 1271
 1272
 1273
 1274
 1275
 1276
 1277
 1278
 1279
 1280
 1281
 1282
 1283
 1284
 1285
 1286
 1287
 1288
 1289
 1290
 1291
 1292
 1293
 1294
 1295
 1296
 1297
 1298
 1299
 1300
 1301
 1302
 1303
 1304
 1305
 1306
 1307
 1308
 1309
 1310
 1311
 1312
 1313
 1314
 1315
 1316
 1317
 1318
 1319
 1320
 1321
 1322
 1323
 1324
 1325
 1326
 1327
 1328
 1329
 1330
 1331
 1332
 1333
 1334
 1335
 1336
 1337
 1338
 1339
 1340
 1341
 1342
 1343
 1344
 1345
 1346
 1347
 1348
 1349
 1350
 1351
 1352
 1353
 1354
 1355
 1356
 1357
 1358
 1359
 1360
 1361
 1362
 1363
 1364
 1365
 1366
 1367
 1368
 1369
 1370
 1371
 1372
 1373
 1374
 1375
 1376
 1377
 1378
 1379
 1380
 1381
 1382
 1383
 1384
 1385
 1386
 1387
 1388
 1389
 1390
 1391
 1392
 1393
 1394
 1395
 1396
 1397
 1398
 1399
 1400
 1401
 1402
 1403
 1404
 1405
 1406
 1407
 1408
 1409
 1410
 1411
 1412
 1413
 1414
 1415
 1416
 1417
 1418
 1419
 1420
 1421
 1422
 1423
 1424
 1425
 1426
 1427
 1428
 1429
 1430
 1431
 1432
 1433
 1434
 1435
 1436
 1437
 1438
 1439
 1440
 1441
 1442
 1443
 1444
 1445
 1446
 1447
 1448
 1449
 1450
 1451
 1452
 1453
 1454
 1455
 1456
 1457
 1458
 1459
 1460
 1461
 1462
 1463
 1464
 1465
 1466
 1467
 1468
 1469
 1470
 1471
 1472
 1473
 1474
 1475
 1476
 1477
 1478
 1479
 1480
 1481
 1482
 1483
 1484
 1485
 1486
 1487
 1488
 1489
 1490
 1491
 1492
 1493
 1494
 1495
 1496
 1497
 1498
 1499
 1500
 1501
 1502
 1503
 1504
 1505
 1506
 1507
 1508
 1509
 1510
 1511
 1512
 1513
 1514
 1515
 1516
 1517
 1518
 1519
 1520
 1521
 1522
 1523
 1524
 1525
 1526
 1527
 1528
 1529
 1530
 1531
 1532
 1533
 1534
 1535
 1536
 1537
 1538
 1539
 1540
 1541
 1542
 1543
 1544
 1545
 1546
 1547
 1548
 1549
 1550
 1551
 1552
 1553
 1554
 1555
 1556
 1557
 1558
 1559
 1560
 1561
 1562
 1563
 1564
 1565
 1566
 1567
 1568
 1569
 1570
 1571
 1572
 1573
 1574
 1575
 1576
 1577
 1578
 1579
 1580
 1581
 1582
 1583
 1584
 1585
 1586
 1587
 1588
 1589
 1590
 1591
 1592
 1593
 1594
 1595
 1596
 1597
 1598
 1599
 1600
 1601
 1602
 1603
 1604
 1605
 1606
 1607
 1608
 1609
 1610
 1611
 1612
 1613
 1614
 1615
 1616
 1617
 1618
 1619
 1620
 1621
 1622
 1623
 1624
 1625
 1626
 1627
 1628
 1629
 1630
 1631
 1632
 1633
 1634
 1635
 1636
 1637
 1638
 1639
 1640
 1641
 1642
 1643
 1644
 1645
 1646
 1647
 1648
 1649
 1650
 1651
 1652
 1653
 1654
 1655
 1656
 1657
 1658
 1659
 1660
 1661
 1662
 1663
 1664
 1665
 1666
 1667
 1668
 1669
 1670
 1671
 1672
 1673
 1674
 1675
 1676
 1677
 1678
 1679
 1680
 1681
 1682
 1683
 1684
 1685
 1686
 1687
 1688
 1689
 1690
 1691
 1692
 1693
 1694
 1695
 1696
 1697
 1698
 1699
 1700
 1701
 1702
 1703
 1704
 1705
 1706
 1707
 1708
 1709
 1710
 1711
 1712
 1713
 1714
 1715
 1716
 1717
 1718
 1719
 1720
 1721
 1722
 1723
 1724
 1725
 1726
 1727
 1728
 1729
 1730
 1731
 1732
 1733
 1734
 1735
 1736
 1737
 1738
 1739
 1740
 1741
 1742
 1743
 1744
 1745
 1746
 1747
 1748
 1749
 1750
 1751
 1752
 1753
 1754
 1755
 1756
 1757
 1758
 1759
 1760
 1761
 1762
 1763
 1764
 1765
 1766
 1767
 1768
 1769
 1770
 1771
 1772
 1773
 1774
 1775
 1776
 1777
 1778
 1779
 1780
 1781
 1782
 1783
 1784
 1785
 1786
 1787
 1788
 1789
 1790
 1791
 1792
 1793
 1794
 1795
 1796
 1797
 1798
 1799
 1800
 1801
 1802
 1803
 1804
 1805
 1806
 1807
 1808
 1809
 1810
 1811
 1812
 1813
 1814
 1815
 1816
 1817
 1818
 1819
 1820
 1821
 1822
 1823
 1824
 1825
 1826
 1827
 1828
 1829
 1830
 1831
 1832
 1833
 1834
 1835
 1836
 1837
 1838
 1839
 1840
 1841
 1842
 1843
 1844
 1845
 1846
 1847
 1848
 1849
 1850
 1851
 1852
 1853
 1854
 1855
 1856
 1857
 1858
 1859
 1860
 1861
 1862
 1863
 1864
 1865
 1866
 1867
 1868
 1869
 1870

432
 433 **Table 4: Process-Accuracy (P-Acc) and Answer-Accuracy (A-Acc)** for CoTs sampled by the base
 434 model (Aya 8B) and RL-tuned derivations. Training on geo-guessing and testing on year-guessing
 435 (and vis versa) constitutes out-of-domain evaluation. The RL training is repeated three times with
 436 different random seeds and mean \pm std is reported. P-Acc and A-Acc are calculated using 200 and
 437 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 MCQ-DrGRPO (ours) | Geo-Guessing | 50.2 \pm 2.5 | 55.6 \pm 0.1 | 56.8 \pm 3.9 | 42.6 \pm 0.2 |
| | | 55.2 \pm 0.6 | 56.0 \pm 0.3 | 58.3 \pm 3.1 | 42.6 \pm 0.1 |
| DrGRPO MCQ-DrGRPO (ours) | Year-Guessing | 57.7 \pm 4.2 | 46.0 \pm 0.2 | 52.7 \pm 1.0 | 52.8 \pm 0.3 |
| | | 59.0 \pm 0.7 | 46.9 \pm 0.3 | 57.5 \pm 2.9 | 53.4 \pm 0.4 |

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

6 RELATED WORK

461 **CoT Process-Correctness** The evaluation of CoT reasoning is primarily focused on answer
 462 correctness (Wei et al., 2022; Wang et al., 2023b; Fu et al., 2023; Liu et al., 2023; DeepSeek-AI et al.,
 463 2025). Because a correct answer does not imply correct reasoning (Wang et al., 2023a), evaluating
 464 the process is of interest for many applications (Singhal et al., 2022; Blair-Stanek et al., 2023;
 465 Macina et al., 2023). Process correctness in compositional reasoning tasks can often be verified by
 466 a parser (Cobbe et al., 2021; Willig et al., 2022; Lyu et al., 2023; Xu et al., 2024), but most natural
 467 language tasks require human annotators (Collins et al., 2022; Zelikman et al., 2022; Uesato et al.,
 468 2023; Mondorf & Plank, 2024). Only recently have studies explored training models (Golovneva
 469 et al., 2023; Prasad et al., 2023) or using LLMs (He et al., 2024; Hao et al., 2024; Bavaresco et al.,
 470 2025) to judge the correctness of reasoning chains. Uesato et al. (2023) show that process reward
 471 models (PRM), learned from human annotations, are improving process correctness at test time. Re-
 472 cently, PRM training moved away from human annotations (Lightman et al., 2024) towards implicit
 473 step-level feedback derived from final answer correctness (Yuan et al., 2025; Wang et al., 2024).

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

7 CONCLUSION

482 We explicitly modeled the ability of an LLM with CoT reasoning to solve a certain multiple-choice
 483 question. To this end, a group of sampled CoTs is used to estimate the probability that the true
 484

486 performance of the LLM exceeds random guessing. We incorporated the estimated solvability of
 487 a question into the objective of an outcome-based reward model and reinforcement learning with
 488 group-relative advantage estimation. Experiments on different base models and datasets showed
 489 improved process accuracy of emitted CoTs, and additionally improved answer accuracy in the
 490 case of RL. Supporting experiments confirmed that answer-correct CoTs from groups with few
 491 correct answers are more likely to be process-incorrect and provide noisy learning signal. The
 492 method does not introduce additional hyperparameter and requires negligible computation overhead.
 493 Taking solvability into account should become the default when using multiple-choice questions for
 494 reasoning training. Although it is not straightforward to adapt our solvability formulation to open-
 495 answer tasks, one possible direction is reformulation to MCQA (Zhang et al., 2025a). Our work
 496 highlights that unsolvable questions are a source of *false positive* CoTs and a method to mitigate
 497 this phenomenon in the MCQA setting.
 498

500 REPRODUCIBILITY

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

504 REFERENCES

505 Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Al-
 506 bert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins,
 507 Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessan-
 508 dro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human
 509 judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce
 510 Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd An-
 511 nual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp.
 512 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-
 513 89176-252-7. URL <https://aclanthology.org/2025.acl-short.20/>.
 514

515 Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statu-
 516 tory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial In-
 517 telligence and Law*, ICAIL ’23, pp. 22–31, New York, NY, USA, 2023. Association for Com-
 518 puting Machinery. ISBN 9798400701979. doi: 10.1145/3594536.3595163. URL <https://doi.org/10.1145/3594536.3595163>.
 519

520 Minghan Chen, Guikun Chen, Wenguan Wang, and Yi Yang. Seed-grpo: Semantic entropy en-
 521 hanced grpo for uncertainty-aware policy optimization, 2025. URL <https://arxiv.org/abs/2505.12346>.
 522

523 Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and
 524 Furu Wei. Reasoning with exploration: An entropy perspective on reinforcement learning for
 525 llms, 2025. URL <https://arxiv.org/abs/2506.14758>.
 526

527 Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser,
 528 Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John
 529 Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL
 530 <https://arxiv.org/abs/2110.14168>.
 531

532 Katherine M. Collins, Catherine Wong, Jiahai Feng, Megan Wei, and Joshua B. Tenenbaum. Struc-
 533 tured, flexible, and robust: benchmarking and improving large language models towards more
 534 human-like behavior in out-of-distribution reasoning tasks, 2022. URL <https://arxiv.org/abs/2205.05718>.
 535

536 Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat
 537 Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason
 538 Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang,
 539 Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün,

540 and Sara Hooker. Aya vision: Advancing the frontier of multilingual multimodality, 2025. URL
 541 <https://arxiv.org/abs/2505.08751>.

542

543 DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu,
 544 Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu,
 545 Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao
 546 Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan,
 547 Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao,
 548 Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,
 549 Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang
 550 Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai
 551 Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,
 552 Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang,
 553 Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang,
 554 Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang,
 555 R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shengfeng
 556 Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing
 557 Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen
 558 Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong
 559 Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu,
 560 Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xi-
 561 aoshua Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia
 562 Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng
 563 Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong
 564 Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong,
 565 Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou,
 566 Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying
 567 Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda
 568 Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu,
 569 Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu
 570 Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforce-
 571 ment learning, 2025. URL <https://arxiv.org/abs/2501.12948>.

572 Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean
 573 Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal,
 574 Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of trans-
 575 formers on compositionality. In *Thirty-seventh Conference on Neural Information Processing
 Systems (NeurIPS)*, 2023. URL <https://openreview.net/forum?id=Fkckkr3ya8>.

576 Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. Chain-of-thought hub: A
 577 continuous effort to measure large language models' reasoning performance, 2023. URL <https://arxiv.org/abs/2305.17306>.

578 Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam
 579 Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step rea-
 580 soning. In *The Eleventh International Conference on Learning Representations*, 2023. URL
 581 <https://openreview.net/forum?id=xYlJRpzZtsY>.

582 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad
 583 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan,
 584 Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Ko-
 585 rennev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava
 586 Spataru, Baptiste Roziere, Bethany Biron, Bin Tang, Bobbie Chern, Charlotte Caucheteux,
 587 Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret,
 588 Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius,
 589 Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary,
 590 Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab
 591 AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco
 592 Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind That-
 593 tai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Kore-
 594 vaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra,

594 Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Ma-
 595 hadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu,
 596 Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jong-
 597 so Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasudevan Alwala,
 598 Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid
 599 El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren
 600 Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin,
 601 Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi,
 602 Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew
 603 Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Ku-
 604 mar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoy-
 605 chev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan
 606 Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan,
 607 Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ra-
 608 mon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Ro-
 609 hit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan
 610 Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell,
 611 Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparth, Sheng
 612 Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer
 613 Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman,
 614 Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mi-
 615 haylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor
 616 Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei
 617 Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang
 618 Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Gold-
 619 schlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning
 620 Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh,
 621 Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria,
 622 Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein,
 623 Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, An-
 624 drew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, An-
 625 nie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,
 626 Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leon-
 627 hardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu
 628 Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Mon-
 629 talvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao
 630 Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia
 631 Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide
 632 Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le,
 633 Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily
 634 Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smo-
 635 thers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni,
 636 Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia
 637 Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan,
 638 Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harri-
 639 son Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj,
 640 Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James
 641 Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-
 642 nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang,
 643 Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Jun-
 644 jie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy
 645 Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang,
 646 Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell,
 647 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

648 Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong,
 649 Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent,
 650 Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar,
 651 Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Ro-
 652 driguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy,
 653 Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin
 654 Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon,
 655 Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ra-
 656 maswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha,
 657 Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal,
 658 Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satter-
 659 field, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj
 660 Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo
 661 Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook
 662 Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Ku-
 663 mar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov,
 664 Wei Li, Wencheng Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiao-
 665 jian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia,
 666 Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao,
 667 Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhao-
 668 duo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL
 669 <https://arxiv.org/abs/2407.21783>.

670 Shibo Hao, Yi Gu, Haotian Luo, Tianyang Liu, Xiyan Shao, Xinyuan Wang, Shuhua Xie, Haodi Ma,
 671 Adithya Samavedhi, Qiyue Gao, Zhen Wang, and Zhitong Hu. LLM reasoners: New evaluation,
 672 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>.

673 Hangfeng He, Hongming Zhang, and Dan Roth. SocREval: Large language models with the
 674 socratic method for reference-free reasoning evaluation. In Kevin Duh, Helena Gomez, and
 675 Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*,
 676 pp. 2736–2764, Mexico City, Mexico, June 2024. Association for Computational Linguistics.
 677 doi: 10.18653/v1/2024.findings-naacl.175. URL <https://aclanthology.org/2024.findings-naacl.175/>.

678 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn
 679 Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset.
 680 In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks
 Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=7Bywt2mQsCe>.

681 Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and
 682 Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to
 683 reproducibility. In *Second Conference on Language Modeling (COLM)*, 2025. URL <https://openreview.net/forum?id=90UrTTxp50>.

684 Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. *arXiv*,
 685 abs/2311.14648, 2024. URL <https://arxiv.org/abs/2311.14648>.

686 Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models
 687 hallucinate. *arXiv*, abs/2509.04664, 2025. URL <https://arxiv.org/abs/2509.04664>.

688 Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 reinforce samples, get a baseline for free!
 689 In *ICLR 2019 Deep Reinforcement Learning meets Structured Prediction Workshop*, 2019.

690 Martha Larson, Mohammad Soleymani, Guillaume Gravier, Bogdan Ionescu, and Gareth J.F. Jones.
 691 The benchmarking initiative for multimedia evaluation: Mediaeval 2016. *IEEE MultiMedia*, 24
 692 (1):93–96, 2017. doi: 10.1109/MMUL.2017.9.

693 Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang,
 694 Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann
 695 Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [<https://huggingface.com>.

702 co/AI-MO/NuminaMath-1.5] (https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

703

704

705 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan
706 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth*
707 *International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=v8L0pN6EOi>.

708

709 Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale gener-
710 ation: Learning to solve and explain algebraic word problems. *ACL*, 2017.

711

712 Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the log-
713 ical reasoning ability of chatgpt and gpt-4, 2023. URL <https://arxiv.org/abs/2304.03439>.

714

715 Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee,
716 and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint*
717 *arXiv:2503.20783*, 2025.

718

719 Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki,
720 and Chris Callison-Burch. Faithful chain-of-thought reasoning. In Jong C. Park, Yuki Arase, Bao-
721 tian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of*
722 *the 13th International Joint Conference on Natural Language Processing and the 3rd Conference*
723 *of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long*
724 *Papers)*, pp. 305–329, Nusa Dua, Bali, November 2023. Association for Computational Linguis-
725 *tics*. doi: 10.18653/v1/2023.ijcnlp-main.20. URL [https://aclanthology.org/2023.ijcnlp-main.20/](https://aclanthology.org/2023.ijcnlp-main.20).

726

727 Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych,
728 and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties
729 grounded in math reasoning problems. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.),
730 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5602–5621, Sin-
731 gapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.
732 *findings-emnlp.372*. URL [https://aclanthology.org/2023.findings-emnlp.372/](https://aclanthology.org/2023.findings-emnlp.372).

733

734 Philipp Mondorf and Barbara Plank. Comparing inferential strategies of humans and large lan-
735 guage models in deductive reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar
736 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Lin-
737 guistics (Volume 1: Long Papers)*, pp. 9370–9402, Bangkok, Thailand, August 2024. Asso-
738 ciation for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.508. URL <https://aclanthology.org/2024.acl-long.508/>.

739

740 Eric Müller, Matthias Springstein, and Ralph Ewerth. “when was this picture taken?” – image
741 date estimation in the wild. In Joemon M Jose, Claudia Hauff, Ismail Sengor Altingovde, Dawei
742 Song, Dyaa Albakour, Stuart Watt, and John Tait (eds.), *Advances in Information Retrieval*, pp.
743 619–625, Cham, 2017. Springer International Publishing. ISBN 978-3-319-56608-5.

744

745 Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David
746 Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Au-
747 gustus Odena. Show your work: Scratchpads for intermediate computation with language models.
748 *arXiv*, abs/2112.00114, 2021. URL <https://arxiv.org/abs/2112.00114>.

749

750 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
751 cia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red
752 Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Moham-
753 mad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher
754 Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brock-
755 man, 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,

756 Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila
 757 Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix,
 758 Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-
 759 son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan
 760 Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hal-
 761 lacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan
 762 Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu,
 763 Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun
 764 Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-
 765 mali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook
 766 Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel
 767 Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen
 768 Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel
 769 Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez,
 770 Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv
 771 Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney,
 772 Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick,
 773 Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel
 774 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Ra-
 775 jeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe,
 776 Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel
 777 Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe
 778 de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny,
 779 Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl,
 780 Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra
 781 Rimbach, Carl Ross, Bob Rotstetd, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders,
 782 Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Sel-
 783 sam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,
 784 Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky,
 785 Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang,
 786 Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-
 787 ston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-
 788 jayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan
 789 Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng,
 790 Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Work-
 791 man, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming
 792 Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao
 793 Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL
 794 <https://arxiv.org/abs/2303.08774>.
 795

796 OpenStreetMap Foundation. Nominatim: Open source geocoding with OpenStreetMap data, 2009.
 797 URL <https://nominatim.org/>.
 798

799 Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale
 800 multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores,
 801 George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Con-
 802 ference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Re-
 803 search*, pp. 248–260. PMLR, 07–08 Apr 2022. URL <https://proceedings.mlr.press/v174/pal22a.html>.
 804

805 Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring
 806 and improving faithfulness of chain-of-thought reasoning. In Yaser Al-Onaizan, Mohit Bansal,
 807 and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP*
 808 2024, pp. 15012–15032, Miami, Florida, USA, November 2024. Association for Computational
 809 Linguistics. doi: 10.18653/v1/2024.findings-emnlp.882. URL <https://aclanthology.org/2024.findings-emnlp.882/>.
 810

811 Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. ReCEval: Evaluating reason-
 812 ing chains via correctness and informativeness. In Houda Bouamor, Juan Pino, and Kalika Bali
 813

810 (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/>.

811

812

813

814

815 Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan
816 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
817 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
818 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,
819 Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang,
820 Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.
821 URL <https://arxiv.org/abs/2412.15115>.

822

823

824

825

826 Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Common-
827 sense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical
828 Methods in Natural Language Processing and the 9th International Joint Conference on Natural
829 Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, 2019.

830

831

832

833

834 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
835 optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL <http://arxiv.org/abs/1707.06347>.

836

837

838

839

840

841 Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang,
842 Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathe-
843 matical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

1081

1082

1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

1129

1130

1131

1132

1133

1134

1135

1136

1137

1138

1139

1140

1141

1142

1143

1144

1145

1146

1147

1148

1149

1150

1151

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187

1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

1237

1238

1239

1240

1241

1242

1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

1296

1297

1298

1299

1300

1301

1302

1303

1304

1305

1306

1307

1308

1309

1310

1311

1312

1313

1314

1315

1316

1317

1318

1319

1320

1321

1322

1323

1324

1325

1326

1327

1328

1329

1330

1331

1332

1333

1334

1335

1336

1337

1338

1339

1340

1341

1342

1343

1344

1345

1346

1347

1348

1349

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

1404

1405

1406

1407

1408

1409

1410

1411

1412

1413

1414

1415

1416

1417

1418

1419

1420

1421

1422

1423

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1436

1437

1438

1439

1440

1441

1442

1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456

1457

1458

1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

1494

1495

1496

1497

1498

1499

1500

1501

1502

1503

1504

1505

1506

1507

1508

1509

1510

1511

1512

1513

1514

1515

1516

1517

1518

1519

1520

1521

1522

1523

1524

1525

1526

1527

1528

1529

1530

1531

1532

1533

1534

1535

1536

1537

1538

1539

1540

1541

1542

1543

1544

1545

1546

1547

1548

1549

1550

1551

1552

1553

1554

1555

1556

1557

1558

1559

1560

1561

1562

1563

1564

1565

1566

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593

1594

1595

1596

1597

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619

1620

1621

1622

1623

1624

1625

1626

1627

1628

1629

1630

1631

1632

1633

1634

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

1674

1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

1723

1724

1725

1726

1727

1728

1729

1730

1731

1732

1733

1734

1735

1736

1737

1738

1739

1740

1741

1742

1743

1744

1745

1746

1747

1748

1749

1750

1751

1752

1753

1754

1755

1756

1757

1758

1759

1760

1761

1762

1763

1764

1765

1766

1767

1768

1769

1770

1771

1772

1773

1774

1775

1776

1777

1778

1779

1780

1781

1782

1783

1784

1785

1786

1787

1788

1789

1790

1791

1792

1793

1794

1795

1796

1797

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

1836

1837

1838

1839

1840

1841

1842

1843

1844

1845

1846

1847

1848

1849

1850

1851

1852

1853

1854

1855

1856

1857

1858

1859

1860

1861

1862

1863

1864

1865

1866

1867

1868

1869

1870

1871

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

1886

1887

1888

1889

1890

1891

1892

1893

1894

1895

1896

1897

1898

1899

1900

1901

1902

1903

1904

1905

1906

1907

1908

1909

1910

1911

1912

1913

1914

1915

1916

1917

1918

1919

1920

1921

1922

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

1944

1945

1946

1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

1998

1999

2000

2001

2002

2003

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014</p

864 Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Yamamoto Siegel, Lisa
 865 Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems
 866 with process-based and outcome-based feedback, 2023. URL [https://openreview.net/](https://openreview.net/forum?id=MND1kmmNy0O)
 867 forum?id=MND1kmmNy0O.

868 Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun.
 869 Towards understanding chain-of-thought prompting: An empirical study of what matters. In
 870 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual*
 871 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2717–
 872 2739, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/
 873 v1/2023.acl-long.153. URL <https://aclanthology.org/2023.acl-long.153/>.

874 Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhi-
 875 fang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In
 876 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meet-*
 877 *ing of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439,
 878 Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/
 879 2024.acl-long.510. URL <https://aclanthology.org/2024.acl-long.510/>.

880 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha
 881 Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language
 882 models. In *The Eleventh International Conference on Learning Representations*, 2023b. URL
 883 <https://openreview.net/forum?id=1PL1NIMMrw>.

884 Zezhong Wang, Xingshan Zeng, Weiwen Liu, Yufei Wang, Liangyou Li, Yasheng Wang, Lifeng
 885 Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Chain-of-probe: Examining the necessity and
 886 accuracy of CoT step-by-step. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings*
 887 *of the Association for Computational Linguistics: NAACL 2025*, pp. 2586–2606, Albuquerque,
 888 New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-
 889 7. doi: 10.18653/v1/2025.findings-naacl.140. URL [https://aclanthology.org/2025.](https://aclanthology.org/2025.findings-naacl.140/)
 890 [findings-naacl.140/](https://aclanthology.org/2025.findings-naacl.140/).

891 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi,
 892 Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language
 893 models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Ad-*
 894 *vances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/forum?id=_VjQ1MeSB_J](https://openreview.net/

 895 <a href=).

896 Xumeng Wen, Zihan Liu, Shun Zheng, Zhijian Xu, Shengyu Ye, Zhirong Wu, Xiao Liang, Yang
 897 Wang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable
 898 rewards implicitly incentivizes correct reasoning in base llms, 2025. URL [https://arxiv.](https://arxiv.org/abs/2506.14245)
 899 [org/abs/2506.14245](https://arxiv.org/abs/2506.14245).

900 Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation mod-
 901 els talk causality?, 2022. URL <https://arxiv.org/abs/2206.10591>.

902 Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful log-
 903 ical reasoning via symbolic chain-of-thought. In Lun-Wei Ku, Andre Martins, and Vivek
 904 Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Compu-*
 905 *tational Linguistics (Volume 1: Long Papers)*, pp. 13326–13365, Bangkok, Thailand, August
 906 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.720. URL
 907 <https://aclanthology.org/2024.acl-long.720/>.

908 Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan
 909 Liu, and Hao Peng. Free process rewards without process labels. In *Forty-second International*
 910 *Conference on Machine Learning*, 2025. URL [https://openreview.net/](https://openreview.net/forum?id=8ThnPfHGM8)
 911 [forum?id=8ThnPfHGM8](https://openreview.net/forum?id=8ThnPfHGM8).

912 Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao
 913 Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond
 914 the base model? In *2nd AI for Math Workshop @ ICML 2025*, 2025. URL [https://openreview.net/forum?id=upehLVgq1b](https://openreview.net/

 915 <a href=).

918 Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STar: Bootstrapping reasoning with
919 reasoning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Ad-*
920 *vances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/](https://openreview.net/forum?id=_3ELRdg2sgI)
921 [forum?id=_3ELRdg2sgI](https://openreview.net/forum?id=_3ELRdg2sgI).

922 Jixiao Zhang and Chunsheng Zuo. Grpo-lead: A difficulty-aware reinforcement learning approach
923 for concise mathematical reasoning in language models. *arXiv preprint arXiv:2504.09696*, 2025.

925 Mengyu Zhang, Xubo Liu, Siyu Ding, Weichong Yin, Yu Sun, Hua Wu, Wenyu Guo, and Ying
926 Zhang. Auditable-choice reframing unlocks rl-based verification for open-ended tasks, 2025a.
927 URL <https://arxiv.org/abs/2511.02463>.

928 Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Edge-grpo: Entropy-driven grpo with
929 guided error correction for advantage diversity, 2025b. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2507.21848)
930 [2507.21848](https://arxiv.org/abs/2507.21848).

931 Ziyin Zhang, Lizhen Xu, Zhaokun Jiang, Hongkun Hao, and Rui Wang. Multiple-choice questions
932 are efficient and robust llm evaluators, 2024.

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972 APPENDIX
973974 A GEO-GUESSING AND YEAR-GUESSING MCQA
975

976 We introduce two multiple-choice question answering (MCQA) datasets for multimodal reasoning.
 977 Both are derived from the YFCC100M dataset (Thomee et al., 2016) that provides 100M images
 978 from Flickr, partially annotated with metadata like tags, title and geolocation. In the following,
 979 we describe the construction of the Geo-Guessing MCQA and Year-Guessing MCQA datasets. For
 980 Geo-Guessing MCQA, we start with MP16 (Larson et al., 2017), a YFCC100M subset of 7M images
 981 that are tagged with geographic coordinates in the form of latitude and longitude. We use Nominatim
 982 (OpenStreetMap Foundation, 2009) to translate the coordinates into a textual description of the
 983 region where an image was taken. We downloaded each image in the highest available resolution,
 984 sorted them by resolution and selected the top 100k instances. These instances are then split into 92k
 985 training, 4k development and 4k test instances. The process for Year-Guessing MCQA is the same
 986 as for geo, except that we start with DEW (Müller et al., 2017), also a subset of YFCC100M, and
 987 extract the year an image was taken from DEW’s additional annotation. They derived the year an
 988 image was taken from user-provided metadata like title, description and tags. An example instance
 989 of Geo-Guessing MCQA is shown in Figure 7 and of Year-Guessing MCQA in Figure 8.
 990

991 A.1 GENERATING ANSWER CHOICES

992 Generating incorrect answer choices (distractors) for MCQA should be handled with care to avoid
 993 exposing the correct answer via subtle bias. For example, generating distractor years via symmetric
 994 error, e.g., $\text{uniform}(\text{year} - \text{distance}, \text{year} + \text{distance})$, let’s a model learn to predict the median
 995 value as correct. On the other hand, to have strong distractors, we need to generate choices that
 996 are close to the correct answer. To this end, we designed an algorithm that is not biased by the
 997 correct value and offers parameters to regulate the maximum distance to the correct value as well
 998 as the minimum distance between answers. The latter is needed to increase distinguishability. The
 999 algorithm is laid out in Algorithm 1 and the parameters for Year-Guessing are set to $n = 2, d =$
 1000 $20, s = 4$. The algorithm first shifts the center in a random direction by a maximum distance of
 1001 $d//2$. Then it samples the distractors uniformly around the shifted center, with a maximum distance
 1002 of $d//2$. The process is repeated until n distractors are generated that all satisfy the minimum
 1003 intra-distance constraint. For Geo-Guessing we modify the algorithm slightly. The correct value
 1004 is a pair of latitude and longitude, the value is shifted by a distance in kilometers and random
 1005 bearing ($\text{uniform}(1, 360)$). Distractors that have the same textual representation (by Nominatim)
 1006 as any of the existing answers are considered equal, even if they have different coordinates. The
 1007 parameters for Geo-Guessing are $n = 3, d = 2000, s = 5$.
 1008

Algorithm 1 Generate Distractor Values with Distance Constraints

1009 **Require:** correct value a , number of distractors n , maximum distance to correct value d , minimum
 1010 distance between answers s
 1011 1: $\text{shifted} \leftarrow \text{RANDOMINT}(a - d//2, a + d//2)$
 1012 2: $\text{answers} \leftarrow \{a\}$
 1013 3: **while** $|\text{answers}| < n + 1$ **do**
 1014 4: $\text{distractor} \leftarrow \text{RANDOMINT}(\text{shifted} - d//2, \text{shifted} + d//2)$
 1015 5: $\text{valid} \leftarrow \text{true}$
 1016 6: **for each** $\text{value} \in \text{answers}$ **do**
 1017 7: **if** $|\text{value} - \text{distractor}| < s$ **then**
 1018 8: $\text{valid} \leftarrow \text{false}$
 1019 9: **end if**
 1020 10: **end for**
 11: **if** valid **then**
 12: $\text{answers} \leftarrow \text{answers} \cup \{\text{distractor}\}$
 13: **end if**
 14: **end while**
 15: **output** answers
 1024

1026
 1027
 1028
 1029
 1030
 1031
 1032
 1033
 1034
 1035
 1036
 1037
 1038
 1039
 1040
 1041
 1042
 1043
 1044
 1045
 1046
 1047
 1048
 1049



1050 **Question:** Can you deduce the country, region and city shown in this photograph?
 1051 **Options:** A. Amsterdam, North Holland, Netherlands B. Scotland, United Kingdom
 1052 C. High Peak, Derbyshire, East Midlands, United Kingdom D. Risør, Agder, Norway
 1053 **Gold Answer:** A

1054 **DrGRPO Chain-of-Thought:**
 1055 This image shows a bustling train station with multiple tracks, cranes, and urban buildings around it. The
 1056 architecture and style of the buildings, along with the presence of trains and trams, suggest a European
 1057 city with a well-developed public transport system. The architecture strongly hints at a city in the
 1058 Netherlands, particularly Amsterdam, known for its canal system and historic buildings. The presence
 1059 of cranes and the layout of the train station are consistent with the infrastructure found in major Dutch
 1060 cities. **Additionally, the signage in the image includes the name "Munsters," which is a local landmark**
 1061 **in Amsterdam.** The surrounding urban landscape and the style of the buildings further support this
 1062 deduction.

1062 **Judge Label:** Incorrect

1063 **MCQ-DrGRPO Chain-of-Thought:**
 1064 This image shows a busy railway station with multiple tracks, a large archway, and urban development in
 1065 the background. The presence of trains, cranes, and a variety of buildings suggests a major transportation
 1066 hub in a densely populated area. The architecture and infrastructure styles are consistent with those
 1067 found in the Netherlands, particularly in Amsterdam. The stations layout and the surrounding buildings
 1068 resemble those found in Amsterdam's central train station and its surrounding areas. Together, these
 1069 elements strongly indicate that the photograph was taken in Amsterdam, in the province of North
 1070 Holland, within the Netherlands.

1070 **Judge Label:** Correct

1071
 1072
 1073 Figure 7: An instance of the Geo-Guessing MCQA dataset with two answer-correct CoTs generated
 1074 by DrGRPO and MCQ-DrGRPO, respectively. The CoTs are evaluated for process-correctness by
 1075 an LLM judge. The text marked in red is a hallucination flagged by the judge.
 1076
 1077
 1078
 1079



1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
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.

1134 Table 5: Hyperparameter for reward model training. Hyperparameter were selected by cross-entropy
 1135 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 |

1145 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

1166
 1167 We list the hyperparameter used for reward model training in Table 5. We ran a minimal grid
 1168 search to find the best hidden dimensions, learning rate, schedule and warmup for the baseline ORM
 1169 and MCQ-ORM. The hyperparameters were chosen based on the lowest loss on the development
 1170 set. The hyperparameter for reinforcement learning experiments are listed in Table 6. They were
 1171 chosen based on initial experiments with DrGRPO and Llama3 1B. The maximum prompt length
 1172 for multimodal Aya is higher because it includes image tokens. Due to resource constraints, we set
 1173 the number of samples to 8 and did not update the image encoder as well as the lower half (16) of
 1174 the LLM layers.

D META EVALUATION

1177
 1178 We use GPT-4.1 (OpenAI et al., 2024) with version *gpt-4.1-2025-04-14* as the judge to assess process
 1179 correctness. To ensure future comparison and reproducibility of our results, we released the
 1180 verbatim CoTs and the full assessment of the judge. The prompt we used is shown in Figure 9.
 1181 We further conduct a meta evaluation that compares the LLM judge with human judgments and its
 1182 ability to detect synthetically corrupted gold CoTs. Golovneva et al. (2023) released a dataset of
 1183 200 model generated CoTs for GSM8k together with human judgments of process correctness. We
 1184 compare these human judgments with those from our LLM judge in Table 7. In 97% of the cases the
 1185 human and LLM judge agree on the process correctness. Four of the six cases in which the human
 1186 and judge disagree are shown in Figure 10, Figure 11, Figure 12 and Figure 13. It is up to the reader
 1187 to decide if the LLM judge is correct or not.

1188 Table 7: Human Meta Evaluation GSM8k. 97% overall.
1189
1190

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

1194
1195 Table 8: Synthetic Meta Evaluation AQuA.
1196
1197

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

1205 to decide whether the concerns by the judge are minor inaccuracies or invalidates the CoT. Next, we
1206 evaluate our LLM judge on a synthetic dataset. To this end, we use the gold rationals of AQuA and
1207 corrupt them by replacing a randomly selected number with a randomly generated number. Table 8
1208 shows the judgment outcomes for the four possible cases. In the first row, we see the most desired
1209 outcome — the gold CoT is judged as correct and the corrupted CoT as incorrect. In only five cases,
1210 the LLM judge does not flag the corrupted CoT as incorrect. Because the LLM judge flags some
1211 gold rationals as incorrect, we manually inspect those CoTs. We find that all of them contain an
1212 error and show an example of an incorrect gold CoT in Figure 14.

1213 Your task is to evaluate whether a reasoning process is fully correct and leads to the ground truth
1214 answer. I will present you a question, its ground truth answer and the reasoning process. Please analyze
1215 the reasoning and tell me if it is fully correct. Return your final assessment as $\boxed{\{yes\}}$ or
1216 $\boxed{\{no\}}$. Remember, the presented answer to the question is the ground truth and you are solely
1217 judging the correctness of the reasoning process.

1218 Question:
1219 <>

1220 Correct Answer:
1221 <>

1222 Reasoning:
1223 <>

1224 Is the reasoning process correct? Think step by step and return your final judgement as $\boxed{\{yes\}}$
1225 or $\boxed{\{no\}}$.
1226

1227 Figure 9: Prompt for LLM judge to assess the process correctness of a CoT.
1228
12291230

E ORM ANSWER-ACCURACY

1231
1232 In addition to the Process-Accuracy (P-Acc) results in Table 1, we report the Answer-Accuracy (A-
1233 Acc) results in Table 9 for the reward model experiments. Here, the ORM baseline outperforms the
1234 MCQ-ORM method. However, in practice, one can use the ORM to select the CoT that is most
1235 likely answer-correct and subsequently the MCQ-ORM to select one of the answer-correct CoTs
1236 that is most likely process-correct.
12371238

F LLM USAGE

1239 We used LLMs to rephrase single sentences, but not full paragraphs. We also used them to help with
1240 the correct formatting of equations and tables.
1241

1242

1243

1244

1245

1246 **Question:** The marching band is ordering new uniforms. Each uniform comes with a hat that costs \$25,
 1247 a jacket that costs three times as much as the hat, and pants that cost the average of the costs of the hat
 1248 and jacket. How much does each uniform cost total?

1249 **Gold Answer:** 150

1250

1251

1252 **Chain-of-Thought:**

1253 Step 1 - The jacket costs $\$25 * 3 = \75 .

1254 Step 2 - The average cost of the hat and jacket is $\$25 + \$75 = \$100$.

1255 Step 3 - The pants cost $\$100 / 2 = \50 .

1256 Step 4 - So the total cost of each uniform is $\$25 + \$75 + \$50 = \150 .

1257 Step 5 - A: 150

1258

1259 **Human Label:** Correct

1260

1261 **Judge Label:** Incorrect

1262 **Judge Reasoning:** Let's analyze the reasoning step by step:

1263 Step 1: The jacket costs $\$25 * 3 = \75 . - This is correct, as the jacket is three times the cost of the hat.

1264 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 $(25 + 75) / 2$.

1265 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.

1266 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.

1267 Step 5: A: 150 - This is the correct answer.

1268 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.

1269 Therefore, the reasoning process contains a misstatement about what the average is, even though the calculation ultimately arrives at the correct answer.

1270 Final assessment: \boxed{\{no\}}

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280 **Table 9: Answer-Accuracy (A-Acc) using different methods of CoT scoring.** The task is to score
 1281 multiple candidate CoTs. The highest scoring CoT is then evaluated for answer-correctness. This
 1282 is done for all development set questions of the three multiple-choice QA datasets. Outcome-
 1283 supervised reward models (ORM) are trained with three different random seeds and mean \pm std is
 1284 reported. Nominal best values are bold.

1285

1286

1287

1288

1289

1290

1291

1292

1293

1294

1295

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

| | 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 \pm 0.1 | 63.8 \pm 0.3 | 86.8 \pm 0.4 | 89.4 \pm 0.1 | 83.7 \pm 0.1 | 98.2 \pm 0.1 |
| MCQ-ORM (ours) | 63.1 \pm 0.1 | 62.7 \pm 0.3 | 85.2 \pm 0.2 | 89.4 \pm 0.1 | 83.6 \pm 0.3 | 98.3 \pm 0.0 |

1296
 1297
 1298
 1299
 1300
 1301
 1302
 1303
 1304
 1305
 1306
 1307
 1308
 1309
 1310 **Question:** Katy makes coffee using teaspoons of sugar and cups of water in the ratio of 7:13. If she used
 1311 a total of 120 teaspoons of sugar and cups of water, calculate the number of teaspoonfuls of sugar she
 1312 used.
 1313 **Gold Answer:** 42
 1314
 1315 **Chain-of-Thought:**
 1316 Step 1 - The total ratio representing the number of teaspoons of sugar and cups of water used is $7+13 =$
 $<<7+13=20>>20$
 1317 Step 2 - The fraction representing the number of teaspoons of sugar used is $7/20$, and since she used 120
 1318 teaspoons of sugar, she used $7/20*120 = <<7/20*120=42.0>>42$ teaspoons of sugar
 1319 Step 3 - A: 42
 1320
 1321 **Human Label:** Correct
 1322
 1323 **Judge Label:** Incorrect
 1324 **Judge Reasoning:** Let's analyze the reasoning step by step:
 1325 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.
 1326 Step 2: The fraction representing the number of teaspoons of sugar used is $7/20$, and since she used 120
 1327 teaspoons of sugar, she used $7/20*120 = 42$ teaspoons of sugar. Here is the issue: The statement says
 1328 "since she used 120 teaspoons of sugar," but the problem says "she used a total of 120 teaspoons of sugar
 1329 and cups of water." That is, 120 is the total of both ingredients, not just sugar. The reasoning should be:
 1330 Number of teaspoons of sugar = $(7/20) \times 120 = 42$.
 1331 The calculation is correct, but the reasoning misstates what the 120 represents. It should say "since she
 1332 used a total of 120 teaspoons of sugar and cups of water," not "since she used 120 teaspoons of sugar."
 1333 Step 3: A: 42 The answer is correct.
 1334 However, because the reasoning misstates what the 120 represents, the reasoning process is not fully
 1335 correct.
 1336 Final assessment: \boxed{\{no\}}

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

1350

1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

1361

1362

1363

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

1368

1369

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 $38 + 2 = <<38+2=40>>40$ seconds.

Step 2 - If Gerald's speed improves by 10%, then his time to run the 400-meter hurdles is reduced by $10/100 * 40 = <<10/100*40=4.0>>4$ seconds.

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

Step 4 - A: 36

1375

1376

Human Label: Correct

1377

1378

Judge Label: Incorrect

1379

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: time = distance / 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{\{no\}}

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403

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

1404
 1405 **Question:** Candice put 80 post-it notes in her purse before she headed out to her job at the coffee shop.
 1406 On her way, she stopped off at the store and purchased a package of Post-it notes; At work, she placed a
 1407 single Post-it note on each of 220 different cups of coffee. If she had 23 post-it notes remaining overall,
 1408 how many Post-it notes were in the package that she purchased?
 1409 **Gold Answer:** 163
 1410
 1411 **Chain-of-Thought:**
 1412 Step 1 - If she placed one post-it note on each of the 220 coffee cups, she used $220/1 = <<220/1=220.0>>220$ post-it notes.
 1413 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.
 1414 Step 3 - If she had 23 post-it notes remaining, then the package contained $140 + 23 = <<140+23=163>>163$ post-it notes.
 1415 Step 4 - A: 163
 1416
 1417 **Human Label:** Correct
 1418
 1419 **Judge Label:** Incorrect
 1420 **Judge Reasoning:** Let's analyze the reasoning step by step:
 1421 **Step 1:** "If she placed one post-it note on each of the 220 coffee cups, she used $220/1 = 220$ post-it
 1422 notes." - This is correct. She used 220 post-it notes.
 1423 **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
 1424 220, she would have a deficit of 140 post-it notes, not that she used 140 from the package. She could not
 1425 have used $220 - 80 = 140$ from the package, because she also had 23 left at the end.
 1426 **Step 3:** "If she had 23 post-it notes remaining, then the package contained $140 + 23 = 163$ post-it
 1427 notes." - This is also incorrect, because the calculation in Step 2 was incorrect. The correct way is to set
 1428 up the equation: $80 + x$ (package) - 220 (used) = 23 (remaining), so $80 + x = 220 + 23$, so $x = 163$.
 1429 **Step 4:** "A: 163" - The answer is correct, but the reasoning to get there is not.
 1430 **Conclusion:** The reasoning process is not fully correct, even though the final answer is correct.
 1431 Final assessment: \boxed{\{no\}}

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

1432
 1433
 1434 **Question:** A trader bought a car at 25% discount on its original price. He sold it at a 40% increase on
 1435 the price he bought it. What percent of profit did he make on the original price?
 1436 **Options:** A. 118 B. 110 C. 105 D. 113 E. 115
 1437 **Gold Answer:** C
 1438 **Gold CoT:** Original price = 100 CP = 75 S = $75*(140/100) = 105$ $100 - 105 = 5\%$ Answer: C
 1439
 1440 **Judge Label:** Incorrect
 1441 **Judge Reasoning:** Let's analyze the reasoning step by step:
 1442 1. Original price = 100
 1443 This is a good assumption for calculation.
 1444 2. CP = 75
 1445 Since the trader bought at a 25% discount, the cost price (CP) is 75.
 1446 3. S = $75*(140/100) = 105$
 1447 He sold it at a 40% increase on the price he bought it, so selling price (S) = $75 \times 1.4 = 105$.
 1448 4. $100 - 105 = 5\%$
 1449 This step is incorrect. The calculation " $100 - 105 = 5\%$ " is not the correct way to find the percent profit
 1450 on the original price.
 1451 The correct way:
 1452 Profit = Selling Price - Original Price = $105 - 100 = 5$ Profit percent on original price = $(\text{Profit} / \text{Original}$
 1453 Price) $\times 100 = (5 / 100) \times 100 = 5\%$
 1454 However, the answer is 5%, which matches option C (105%), but the reasoning is not clearly or correctly
 1455 explained, especially in the last step.
 1456 Conclusion: The reasoning process contains a calculation error in the last step and does not clearly explain
 1457 how the percent profit is calculated on the original price. Therefore, the reasoning is not fully correct.
 1458 \boxed{\{no\}}

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