CoRA: Confidence–Rationale Alignment for Reliable Chain-of-Thought Reasoning

Anonymous ACL submission

Abstract

Chain-of-thought (CoT) reasoning can improve LLM performance, but high answer confidence may be misleading when the accompanying CoT rationale is plausible yet incomplete or poorly supported. We study confidence–rationale alignment: whether a model's confidence in its committed answer is justified by its generated rationale. We introduce a GRPO-based reinforcement learning framework that jointly rewards answer correctness, committed-answer probability, and rubric-based rationale support, where the rubric assesses grounding, coherence, task match, and connection to the selected answer without revealing the gold answer to the judge. Across MedQA, MathQA, and OpenBookQA using three open-weight LLMs, our method reduces the confidence–rationale alignment error by up to 26.51% compared with untuned checkpoints, SFT, and correctness-only GRPO, while maintaining competitive accuracy and often improving calibration. These results show that reliable CoT reasoning requires not only confident answers, but rationales that substantively support them.

1 Introduction

Chain-of-thought (CoT) reasoning has been shown to improve large language model (LLM) performance on arithmetic, commonsense, symbolic, and other reasoning tasks (Wei et al., 2022; Kojima et al., 2022). In this paper, we use *rationale* to refer to the generated CoT explanation that accompanies the final answer. In many high-stakes user-facing settings, the reliability of a response is judged not only by whether the final answer is correct, but also by whether the model appears confident and whether its rationale justifies the answer. This motivates *confidence–rationale alignment*: the extent to which a model's confidence in its committed answer is justified by the rationale it generates.

This problem is important because a model may produce a fluent rationale, choose an incorrect answer, and still assign high confidence to that answer. Such failures can be subtle: when an answer is accompanied by both a persuasive rationale and a strong confidence signal, users may find the error difficult to recognize or correct. Prior work in human-AI interaction shows that rationales and explanations can strongly shape user reliance on model outputs, sometimes increasing reliance even when the response is incorrect (Vasconcelos et al., 2023; Steyvers et al., 2025; Kim et al., 2025). A trustworthy reasoning model should therefore be confident only when its rationale adequately supports the answer it commits to.

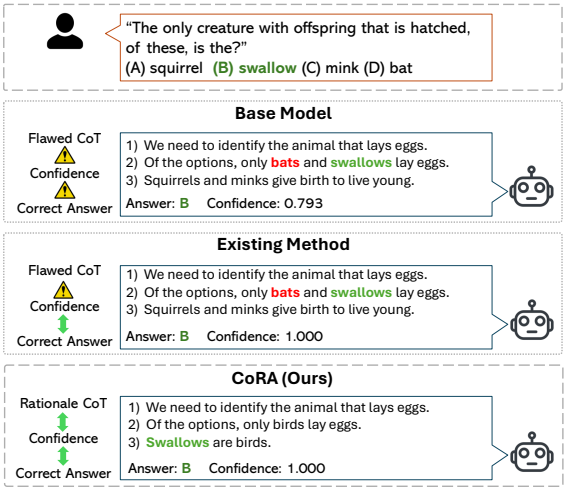

Figure 1: An example of confidence–rationale misalignment. Both the base model and a correctness-oriented method choose the correct answer with near-perfect confidence but include an unsupported claim that bats lay eggs. Our method produces a rationale that better supports the committed answer by connecting the selected option, swallow, to hatched offspring.

Existing work addresses related parts of this problem, but not their interaction. CoT prompting, self-consistency, and tree-structured search improve reasoning accuracy by eliciting or searching over intermediate reasoning traces (Wei et al., 2022;

Wang et al., 2023; Yao et al., 2023). However, generated rationales may not faithfully reflect the factors that determine the model's answer (Turpin et al., 2023; Lanham et al., 2023; Paul et al., 2024; Tutek et al., 2025). Separately, calibration methods aim to make answer confidence reflect empirical correctness (Guo et al., 2017; Zhao et al., 2021; Kadavath et al., 2022; Tian et al., 2023; Xie et al., 2024; Shen et al., 2024). Yet they typically evaluate output probabilities in aggregate and do not assess whether an individual rationale supports the answer at the matched confidence level. This distinction matters in human-facing systems, where confidence and explanation signals can affect user reliance and may not be interpreted uniformly across users (Lin et al., 2022; Tian et al., 2023; Steyvers et al., 2025; Kim et al., 2025).

To address this gap, we introduce a confidence–rationale alignment framework, CoRA, for multiple-choice reasoning. CoRA has two components. First, we use a structured LLM-as-judge rubric to assess whether a rationale is grounded, coherent, task-matched, and properly connected to the model's committed answer without exposing the gold answer (Zheng et al., 2023; Liu et al., 2023; Kim et al., 2024). Second, we optimize the LLM with a Group Relative Policy Optimization (GRPO)-based reward (Shao et al., 2024; DeepSeek-AI, 2025) that combines answer correctness, rationale-support quality, and committed-answer confidence. Unlike correctness-only GRPO, this objective encourages the model to ground its confidence in the generated rationale rather than treating confidence as an isolated scalar.

We evaluate CoRA on three benchmark datasets, MedQA, MathQA, and OpenBookQA (Jin et al., 2021; Amini et al., 2019; Mihaylov et al., 2018), across three open-weight models. We measure answer accuracy, committed-answer probability calibration using expected calibration error (ECE) and Brier score (Brier, 1950), and confidence–rationale mismatch, which captures cases where confidence exceeds rationale-support quality. Empirically, CoRA most consistently reduces unsupported overconfidence, with the clearest gains on MathQA, while maintaining competitive accuracy.

Our contributions are fourfold. (1) We formulate *confidence–rationale alignment* as a reliability problem for reasoning LLMs, requiring committed-answer confidence to be justified by the generated rationale. (2) We design a structured LLM-as-judge rubric that evaluates rationale support for the model's selected answer without exposing the gold answer. (3) We propose a GRPO-based training framework that combines answer correctness, rationale-support quality, and committed-answer confidence to reduce unsupported overconfidence. (4) We evaluate CoRA on three benchmark datasets and three open-weight models, showing reduced confidence–rationale error in most settings while maintaining competitive accuracy; we further introduce a downstream correctness-prediction task showing that CoRA can make generated reasoning traces more diagnostically informative.

2 Related Work

2.1 CoT Reasoning and Rationale Faithfulness

CoT prompting improves LLM reasoning by eliciting intermediate reasoning steps before a final answer (Wei et al., 2022). Follow-up methods such as zero-shot CoT, self-consistency, Tree of Thoughts, and STaR further show that generated reasoning traces can improve task accuracy through prompting, sampling, search, or bootstrapping from model-generated rationales (Kojima et al., 2022; Wang et al., 2023; Yao et al., 2023; Zelikman et al., 2022). These methods demonstrate the practical value of intermediate reasoning traces, but they primarily optimize or evaluate final-answer performance.

Improved reasoning performance, however, does not guarantee that a generated rationale faithfully explains the model's prediction. Prior work shows that CoT rationales can omit factors that influence model outputs, rationalize biased or incorrect answers, or fail to causally determine the final answer (Turpin et al., 2023; Lanham et al., 2023). More recent work uses causal mediation, process verification, or unlearning-based interventions to test whether reasoning steps influence or justify final predictions (Lightman et al., 2023; Paul et al., 2024; Tutek et al., 2025). Our work builds on this reliability concern, but focuses on a different question: whether the model's confidence in its committed answer is justified by the rationale it presents.

2.2 Confidence Calibration and Confidence–Quality Alignment

Confidence calibration aims to ensure that predictive confidence reflects empirical correctness. Classical calibration work shows that neural networks can be poorly calibrated and that post-hoc methods such as temperature scaling can improve

probability estimates (Guo et al., 2017). In LLMs, prior work studies prompt-induced calibration errors, whether models can assess the correctness of their own answers, and how models express uncertainty in probabilities or words (Zhao et al., 2021; Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023). Other work proposes post-hoc or auxiliary calibration methods and surveys broader confidence estimation techniques for LLMs (Xie et al., 2024; Shen et al., 2024; Geng et al., 2024).

Recent work has also begun to connect confidence with response quality. CONQORD aligns verbalized confidence with response quality using reinforcement learning (Tao et al., 2024), while CoT-UQ and CER study whether CoT or confidence signals can improve uncertainty quantification and reasoning behavior (Zhang and Zhang, 2025; Razghandi et al., 2025). These studies are closely related to our motivation, but our setting differs in two ways: 1) we use committed-answer probability rather than verbalized confidence; 2) we explicitly evaluate whether confidence is supported by the rationale, yielding an instance-level alignment signal beyond aggregate calibration.

2.3 Rationale Evaluation, LLM-as-Judge, and RL Training

Our rationale-support rubric is grounded in prior work on rationales, faithfulness, and rubric-based evaluation. These studies investigate how models can provide textual evidence for their decisions (Lei et al., 2016) and provide a benchmark for evaluating rationalized NLP models (DeYoung et al., 2020). Jacovi and Goldberg (2020) further emphasizes that explanations should be evaluated by their relationship to model predictions, not only by their surface plausibility. These ideas motivate our focus on answer-support quality: whether a rationale uses relevant evidence, follows coherent inference steps, and bridges to the model's selected answer.

LLM-as-judge evaluation provides a scalable way to assess open-ended outputs, but prior work also shows that judge behavior can be sensitive to rubric design and may exhibit biases (Zheng et al., 2023). G-Eval and Prometheus demonstrate that structured evaluation prompts and fine-grained rubrics can improve the consistency and usefulness of model-based evaluation (Liu et al., 2023; Kim et al., 2024). Following this line of work, our judge is constrained by a structured rubric and is not given the gold answer, so that it evaluates whether the rationale supports the model's own committed answer rather than whether the answer is correct.

Our optimization method is related to reinforcement learning for LLMs. PPO-style optimization has been widely used in LLM alignment (Schulman et al., 2017; Ouyang et al., 2022), and recent reasoning-oriented RL methods such as DeepSeek-Math and DeepSeek-R1 show that reinforcement learning can improve mathematical and general reasoning behavior (Shao et al., 2024; DeepSeek-AI, 2025). In contrast to correctness-only reinforcement learning, our reward incorporates answer correctness, rationale-support quality, and committed-answer confidence. This connects reasoning supervision with calibration by training models to reduce unsupported overconfidence rather than only maximizing final-answer correctness.

3 Method

We develop a reinforcement learning framework for confidence–rationale alignment in multiple-choice reasoning, as shown in Figure 2.

3.1 Rubric for Assessing Rationale Quality

We define rationale quality as the extent to which a generated rationale provides coherent, grounded, and task-appropriate support for the model's committed answer. We estimate this quality using an LLM-as-judge that takes as input the dataset name, question, answer options, optional auxiliary background knowledge, generated rationale, and committed answer. The auxiliary background knowledge does not include the gold option label, official solution, official explanation, or gold rationale. The judge evaluates whether the rationale supports the committed answer, not whether the committed answer is correct. We treat this judge score as a scalable proxy for rationale-support quality rather than as a human gold standard. To reduce label leakage, the judge is constrained by a structured rubric and is not given the gold answer.

The rubric is designed based on criteria that recur in prior work, which emphasizes 1) whether explanations identify relevant evidence and support model predictions (Lei et al., 2016; DeYoung et al., 2020; Jacovi and Goldberg, 2020); 2) CoT faithfulness work highlights the need to assess whether reasoning steps coherently justify final predictions (Turpin et al., 2023; Lanham et al., 2023; Paul et al., 2024); and 3) LLM-as-judge work motivates explicit, fine-grained rubrics to reduce ambiguity in open-ended evaluation (Zheng et al., 2023; Liu

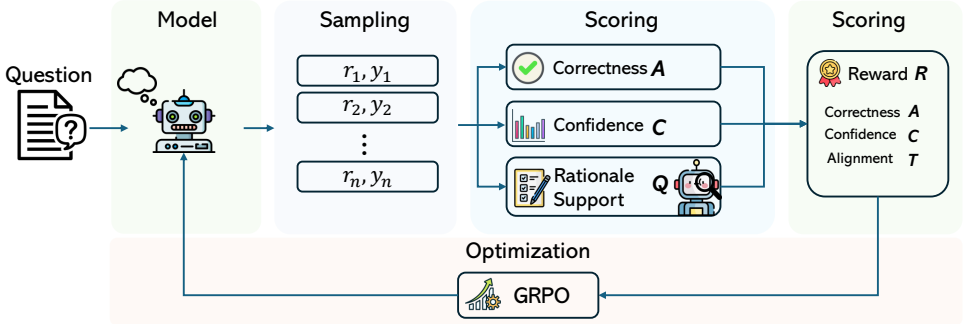

Figure 2: **An overview of the CoRA framework.** Given multiple-choice questions, the policy model samples a group of responses, each consisting of a rationale and a committed answer. For each response, an LLM judge estimates rationale-support quality Q, while the policy model provides committed-answer confidence C. We first combine Q and C into an alignment score T, and then use T, correctness A, and confidence C to compute the final reward R for GRPO optimization. The reward favors well-supported confident answers and penalizes unsupported overconfidence.

Group	Axis	Description
General	G_1	Format validity
	G_2	Task understanding
	G_3	Evidence grounding
	G_4	Inference coherence
	G_5	Answer bridge
	G_6	Structure
OpenBookQA	O_1	Science fact correctness
	O_2	Scenario application
MedQA	Med_1	Medical fact correctness
	Med_2	Clinical application
MathQA	Math_1	Computation correctness
	Math_2	Quantity setup

Table 1: A modular rationale-support rubric. General axes are shared across datasets; each dataset adds two task-specific axes.

et al., 2023; Kim et al., 2024). Based on these principles, our general axes cover format validity, task understanding, evidence grounding, inference coherence, answer bridging, and structure. The task-specific axes add dataset-relevant requirements, such as computation and quantity setup for MathQA, clinical application for MedQA, and science-fact application for OpenBookQA.

The rubric provides an operational measure of answer-support quality. It evaluates whether a generated reasoning trace supports the model's committed answer through grounded, coherent, and task-matched reasoning. To reduce judge-specific bias, we evaluate alignment with three independent judge models and report whether the observed trends remain consistent across judges.

Each axis is assigned a categorical label and mapped to a scalar value in $[0, 1]$. Let Q_{gen} denote the aggregate score over the six general axes, and Q_{task} denote the aggregate score over the two task-specific axes for the corresponding dataset. The final rationale-support score is:

$$Q = w_{\text{gen}} Q_{\text{gen}} + w_{\text{task}} Q_{\text{task}} - P, \qquad (1)$$

where P denotes a penalty for severe structural failure or unsupported reasoning. The resulting score is clipped to $[0, 1]$. The format axis G_1 acts as a gate: if the response lacks a valid parseable answer, Q is set to zero.

3.2 Confidence–Rationale Alignment Reward

We design a reward function that jointly considers answer correctness, rationale-support quality, and committed-answer confidence.

Committed-answer confidence. Let $\mathcal{Y}$ denote the set of answer options and let $y \in \mathcal{Y}$ be the parsed answer generated by the model. We compute the committed-answer confidence C by normalizing model scores over candidate answer labels:

$$C = \frac{\exp(s_y)}{\sum_{\ell \in \mathcal{Y}} \exp(s_\ell)}, \qquad (2)$$

where s_ℓ is the conditional log-probability score of candidate label ℓ under the exact generated prefix up to, but excluding, the final answer label. Thus, $C \in [0, 1]$ measures the normalized probability assigned to the model's committed answer. We use option-label probabilities rather than verbalized confidence. Details are described in Appendix A.1.

Alignment score. A desirable response should maintain consistency between confidence and rationale support: higher confidence should be ac-

companied by stronger rationale support, while weakly supported rationales should receive lower confidence. We operationalize this criterion with a confidence–rationale alignment score T as follows:

$$T = \text{clip}\big(\alpha Q + (1-\alpha)C - \beta_{\text{over}} \max(C - Q, 0) - \beta_{\text{under}} \max(Q - C, 0), 0, 1\big). \quad (3)$$

The first two terms combine rationale support and committed-answer confidence. The penalty terms capture the extent of the confidence–rationale mismatch: $\max(C - Q, 0)$ penalizes confidence that exceeds rationale support, while $\max(Q - C, 0)$ penalizes confidence that falls below rationale support. We use asymmetric penalties to prioritize reducing unsupported high-confidence answers.

Overall reward. The final reward combines correctness and confidence–rationale alignment:

$$R = A(1 + \lambda T) - \gamma(1 - A)C - \beta \max(C - Q, 0). \quad (4)$$

Here, $A \in \{0, 1\}$ is a binary answer-correctness signal, while Q, C, and T are continuous scores in $[0, 1]$. The first term rewards correct answers and those with stronger confidence–rationale alignment. The second term penalizes confident wrong answers, while the third term penalizes confidence that exceeds rationale support, regardless of answer correctness. This reward encourages the model to produce accurate answers, generate rationales that support those answers, and calibrate its confidence according to the quality of its reasoning.

3.3 Optimization

We optimize the policy model using GRPO. For each input question, the policy samples a group of candidate completions. Each completion is parsed into a rationale and a committed answer, scored for answer correctness, evaluated by the LLM judge for rationale-support quality, and assigned a committed-answer confidence score. The resulting reward in Equation 4 is applied to compute group-relative advantages within the sampled response group. The policy is then updated to increase the likelihood of responses with higher relative rewards while regularizing against large deviations from the reference policy. This allows the model to learn from multiple signals simultaneously.

3.4 Correctness Prediction from CoT Reasoning Linguistic Features

As an auxiliary diagnostic of user-facing reasoning transparency, we test whether surface linguistic features of generated reasoning traces contain information about answer correctness. Given a generated reasoning trace r_i, we extract a feature vector $\phi(r_i)$ and train a logistic regression classifier to predict the binary correctness label. We use stratified cross-validation and report out-of-fold predicted probabilities. Feature standardization is performed within each training fold to avoid information leakage. The full feature set is listed in Appendix B Table 5.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate CoRA on three multiple-choice reasoning benchmarks spanning mathematical, medical, and scientific reasoning.

- **MathQA** (Amini et al., 2019): A mathematical reasoning benchmark consisting of multiple-choice word problems that require quantity identification, symbolic reasoning, and arithmetic computation.
- **MedQA** (Jin et al., 2021): A medical question-answering benchmark derived from medical licensing exams. It evaluates clinical and biomedical reasoning over patient-centered and factual medical scenarios.
- **OpenBookQA** (Mihaylov et al., 2018): An elementary science question-answering benchmark that requires using scientific facts and commonsense knowledge for selecting the correct answer.

Models. We use three open-weight models as policy models: Ministral 3 8B Reasoning (Mistral AI, 2025), Qwen2.5 7B (Qwen Team, 2024), and Gemma 2 9B Instruct (Gemma Team, 2024). These models cover different model families and instruction-tuning settings. We use the same prompting format and evaluation protocol across all policy models and datasets.

For rubric-based training, we use GPT-OSS 20B as the judge model. The judge is provided with task-relevant auxiliary background knowledge, but not the gold option label, official solution, official explanation, or gold rationale. This setup approximates a stronger evaluator with access to external

evidence, reducing the chance that the judge rewards or penalizes a reasoning trace due to missing background knowledge.

For evaluation, we use three judge models, GPT-OSS 20B (OpenAI, 2025), Llama 3.1 70B Instruct (Llama Team, AI @ Meta, 2024), and Qwen 3 32B (Yang et al., 2025) to measure rationale-support quality and confidence–rationale alignment. The evaluation judges are also provided with the same auxiliary background knowledge used by the training judge, under the same no-gold-information constraint. This design allows us to assess whether the observed alignment trends remain stable across different judge models rather than depending on a single evaluator.

Baselines. We compare CoRA with three training baselines: 1) **Base**: The original model without task-specific fine-tuning; 2) **SFT**: Supervised fine-tuning on task examples containing rationales and final answers; 3) **GRPO**: GRPO training with a rule-based correctness reward, where the reward is determined by whether the parsed answer matches the gold label. It uses the same optimization setup as CoRA, but replaces the confidence–rationale reward with a correctness-only reward.

4.2 Evaluation Metrics

To evaluate whether CoRA improves both task performance and confidence–rationale alignment, we employ accuracy and calibration metrics, together with rationale-centered alignment diagnostics.

- **Accuracy**: It measures whether the parsed answer matches the gold label. We use the final parsed option as the model prediction.
- **Expected Calibration Error (ECE)** (Guo et al., 2017): ECE measures the discrepancy between committed-answer confidence and empirical correctness. We compute ECE using 10 equal-width confidence bins.
- **Brier Score** (Brier, 1950): It is a proper scoring rule for probabilistic calibration. For each example, it computes the squared difference between committed-answer confidence and binary correctness.
- **Rationale-Support Quality**: We evaluate it using a rubric-based judge, which assigns each generated reasoning trace a score $Q \in [0, 1]$. It measures whether the reasoning trace is grounded, coherent, task-matched, and supportive of the committed answer.
- **Confidence–Rationale Alignment Error**: We evaluate confidence–rationale alignment using the average absolute gap between committed-answer confidence and rationale-support quality:

$$E_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} |C_i - Q_i|. \tag{5}$$

  Lower E_{align} indicates stronger alignment between confidence and rationale support.

4.3 Implementation Details

For each dataset, we use 200 samples for training and 500 held-out samples for evaluation. For SFT, the 200 training examples are selected from instances that the corresponding base model answers correctly. For GRPO and CoRA, we randomly sample 200 examples from the training set. The same evaluation subset is used across all methods for a controlled comparison. During GRPO-based training, we sample 8 completions per question with a sampling temperature of 0.8.

5 Results

5.1 Accuracy and Calibration Scores.

We conduct experiments to evaluate the effectiveness of CoRA on MathQA, MedQA, and OpenBookQA across three foundation models. Table 2 reports answer accuracy, Expected Calibration Error (ECE), and Brier score.

We observe that CoRA generally improves calibration while preserving competitive task performance. In particular, CoRA achieves the best Brier score in most model–dataset settings, including all three datasets for Gemma 2 9B Instruct, MathQA and OpenBookQA for Ministral 3 8B, and MathQA and OpenBookQA for Qwen2.5 7B. This suggests that incorporating confidence–rationale alignment into reinforcement learning improves the probabilistic reliability of the committed answers.

CoRA also maintains relatively strong answer accuracy. On MathQA, CoRA obtains the best or tied-best accuracy for Ministral 3 8B and Gemma 2 9B Instruct, while remaining competitive with GRPO on Qwen2.5 7B. On OpenBookQA, CoRA achieves the best accuracy for Ministral 3 8B and Qwen2.5 7B, and the second-best accuracy for Gemma 2 9B Instruct. These results indicate

Table 2: **Results on accuracy and calibration scores on datasets and models.** Accuracy, ECE, and Brier score are reported as percentages. Higher accuracy is better, while lower ECE and Brier score indicate better calibration. Best values are bolded and second-best values are underlined within each model–dataset block.

		Ministral 3 8B			Qwen2.5 7B			Gemma 2 9B Instruct		
Dataset	**Method**	**Acc ↑**	**ECE ↓**	**Brier ↓**	**Acc ↑**	**ECE ↓**	**Brier ↓**	**Acc ↑**	**ECE ↓**	**Brier ↓**
MathQA	Base	79.40	8.35	26.43	72.00	11.55	37.69	58.80	32.21	68.08
	SFT	**80.20**	12.40	29.51	72.80	17.55	41.69	61.20	34.43	70.67
	GRPO	79.80	7.88	25.86	**77.20**	13.94	35.50	61.80	29.90	64.70
	CoRA (Ours)	**80.20**	**6.58**	**22.66**	76.40	**11.09**	**32.83**	**63.40**	**25.60**	**58.27**
MedQA	Base	65.00	31.56	65.73	46.60	42.94	93.13	51.80	46.86	93.86
	SFT	65.20	33.06	67.83	50.40	44.45	91.39	52.00	47.58	94.83
	GRPO	**66.80**	**30.37**	**62.66**	**54.40**	**35.17**	**76.74**	54.60	43.28	87.56
	CoRA (Ours)	64.60	31.91	65.41	52.00	39.72	84.79	**55.60**	**42.22**	**86.48**
OpenBookQA	Base	83.40	13.94	30.50	80.80	**10.65**	30.45	78.80	19.16	38.77
	SFT	87.20	**11.54**	24.23	82.80	12.71	29.74	80.60	19.05	37.86
	GRPO	86.40	**11.54**	24.78	82.60	10.73	27.80	**84.20**	14.52	30.23
	CoRA (Ours)	**87.40**	11.72	**23.96**	**86.00**	11.89	**25.57**	84.00	**14.14**	**30.21**

Table 3: **Rationale-support quality and confidence–rationale alignment results evaluated by Llama 3.1 70B Instruct.** Q measures the quality of the generated reasoning trace, while E_{align} denotes the average confidence–rationale alignment error. ↑ indicates higher is better, and ↓ indicates lower is better. All values are reported as percentages.

		Ministral 3 8B		Qwen2.5 7B		Gemma 2 9B	
Dataset	**Method**	$Q\uparrow$	$E_{\text{align}}\downarrow$	$Q\uparrow$	$E_{\text{align}}\downarrow$	$Q\uparrow$	$E_{\text{align}}\downarrow$
MathQA	Base	97.56	4.56	86.55	14.35	82.75	16.70
	SFT	96.75	3.92	87.52	12.88	84.01	15.89
	GRPO	97.12	4.55	87.82	10.28	83.86	15.73
	CoRA (Ours)	**98.18**	**1.65**	**91.57**	**9.53**	**89.96**	**11.28**
MedQA	Base	94.90	6.09	73.42	26.20	81.47	18.31
	SFT	94.88	5.35	77.67	21.76	82.59	17.14
	GRPO	**95.65**	5.79	74.77	22.31	82.63	16.91
	CoRA (Ours)	94.99	**5.08**	**79.87**	**20.29**	**94.47**	**6.02**
OpenBookQA	Base	95.34	5.28	91.94	10.44	90.38	9.02
	SFT	**97.58**	**2.58**	91.92	8.48	90.54	9.06
	GRPO	95.91	4.59	90.90	9.46	92.14	7.59
	CoRA (Ours)	96.09	3.66	**94.65**	**5.65**	**92.60**	**7.29**

that optimizing for confidence–rationale alignment does not require sacrificing task accuracy.

However, CoRA is not uniformly best in every setting. On MedQA, GRPO achieves the strongest calibration for Ministral 3 8B and Qwen2.5 7B, while CoRA performs best for Gemma 2 9B Instruct. This suggests that the benefit of confidence–rationale alignment depends on the interaction between model family and domain. Nevertheless, CoRA remains consistently competitive and often improves calibration-sensitive metrics.

5.2 Rationale-Support Quality and Alignment

Table 3 reports reasoning quality Q and alignment error E_{align} evaluated by Llama 3.1 70B Instruct. CoRA shows the clearest gains on MathQA, achieving the highest Q and lowest E_{align} across all three policy models. This indicates that CoRA improves both the quality of mathematical reasoning traces and the consistency between rationale support and committed-answer confidence.

On MedQA, CoRA also performs strongly. It obtains the lowest E_{align} across all three models and the highest Q for Qwen2.5 7B and Gemma 2 9B, while GRPO achieves the highest Q for Ministral 3 8B. On OpenBookQA, CoRA achieves the best results for Qwen2.5 7B and Gemma 2 9B, while SFT performs best for Ministral 3 8B.

Overall, the results show that CoRA generally improves confidence–rationale alignment and reasoning-support quality, with especially consistent gains on MathQA.

To reduce the impact of judge-specific bias in LLM-based evaluation, we further evaluate Q and E_{align} using two additional judge models: GPT-OSS 20B and Qwen 3 32B. These judges follow the same rubric and receive the same auxiliary background knowledge as the main evaluator, under the same no-gold-information constraint. The additional results are reported in Appendix C, from Table 6 to Table 7. Across judges, the overall trends remain broadly consistent, suggesting that the observed improvements are not an artifact of a single evaluator.

Table 4: **Prediction Results from CoT Reasoning Linguistic Features.** ↑ indicates higher is better. Best values are bolded and second-best values are underlined within each model–dataset block.

		Ministral 3 8B	Qwen2.5 7B	Gemma 2 9B
Dataset	**Method**	**AUROC ↑**	**AUROC ↑**	**AUROC ↑**
MathQA	Base	0.820	0.665	0.694
	SFT	0.787	0.651	0.662
	GRPO	0.833	0.592	0.670
	CoRA (Ours)	**0.846**	**0.741**	**0.695**
MedQA	Base	0.588	0.552	0.537
	SFT	0.578	0.555	0.538
	GRPO	**0.613**	0.539	0.503
	CoRA (Ours)	0.572	**0.591**	**0.554**
OpenBookQA	Base	0.610	0.495	**0.560**
	SFT	0.567	0.525	0.479
	GRPO	0.568	0.542	0.489
	CoRA (Ours)	**0.627**	**0.581**	0.503

5.3 Correctness Prediction from CoT Reasoning Linguistic Features

Table 4 evaluates whether surface linguistic features in generated reasoning traces can predict answer correctness. Overall, the results show that linguistic cues provide a useful but limited signal. Many AUROC values on MedQA and OpenBookQA remain close to the weak-discrimination range of 0.50–0.60, indicating that surface linguistic features alone are insufficient for reliable correctness prediction in these settings.

The clearest signal appears on MathQA. Across all models, MathQA obtains substantially higher AUROC than other datasets, and CoRA achieves the best AUROC for every model. For example, on Qwen2.5, CoRA improves linguistic-only AUROC from 0.592 under GRPO to 0.741. This suggests that mathematical reasoning traces contain stronger observable correctness signals, which CoRA further enhances.

The results on MedQA and OpenBookQA are more mixed. CoRA achieves the best AUROC for Qwen2.5 and Gemma 2 on MedQA, and for Ministral 3 and Qwen2.5 on OpenBookQA, but several scores remain near 0.55. This suggests that correctness in medical and science QA may depend more on latent domain knowledge than on shallow linguistic patterns alone. The main takeaway is that CoRA can make generated reasoning traces more diagnostically informative, especially on MathQA.

6 Discussion and Conclusion

The results suggest that confidence–rationale alignment is most useful as a reliability-oriented objective rather than as a direct substitute for correctness-oriented training. CoRA most consistently reduces unsupported overconfidence, as measured by the alignment error E_{align}, while maintaining competitive accuracy. This distinction is important because a model can be accurate or calibrated on average and still produce individual responses whose confidence is not supported by the rationale it presents. Confidence–rationale alignment therefore provides an instance-level reliability signal that complements conventional accuracy and calibration metrics.

The main trade-off is that CoRA improves rationale-centered alignment more consistently than it improves all conventional metrics. Correctness-only GRPO rewards the final answer, whereas CoRA also penalizes cases in which confidence exceeds rationale support. As a result, CoRA may prefer responses that are better supported or less overconfident, even when a correctness-only objective would favor more aggressive predictions. This behavior is desirable when the goal is to reduce unsupported high-confidence reasoning, but it also means that confidence–rationale alignment should be viewed as complementary to answer accuracy, not as a guaranteed way to maximize it.

The domain differences suggest that the value of rationale-support supervision depends on how reliably rationale quality can be assessed. MathQA provides relatively explicit computational and logical structure, which may make rationale-support judgments more stable. MedQA requires specialized domain knowledge and may introduce greater ambiguity for both the policy model and the judge. This helps explain why the gains are broadest on MathQA and more variable on MedQA.

Evaluations under Llama 3.1 70B Instruct, GPT-OSS 20B, and Qwen 3 32B further suggest that the observed improvements are not solely an artifact of the training judge. The fact that CoRA often reduces the alignment error under all judges provides evidence that the method improves a broader rationale-support signal rather than only matching the preferences of a single judge. Meanwhile, disagreements between judges indicate that rationale-support evaluation remains imperfect and should not be treated as a substitute for human assessment.

Overall, our findings support treating answer correctness, confidence, and rationale support as coupled reliability signals. A reasoning model should not merely be accurate or calibrated on average; it should assign high confidence only when its rationale adequately supports its answer.

Limitations

There are several limitations to this work that we believe serve as opportunities for further investigation and methodological advancement. First, our rationale-support score depends on LLM judges, whose judgments may contain biases or inconsistencies despite the use of a structured rubric and evaluation across multiple judge models. Future work should compare judge scores with expert and non-expert human assessments of rationale support. Second, our current experiments use relatively small training subsets, so stronger trends may require larger-scale training and multiple random seeds. Third, although the alignment error E_{align} captures unsupported overconfidence, it does not fully measure all forms of rationale faithfulness. A rationale may appear supportive according to the rubric while still not reflecting the model's internal decision process. Finally, our current evaluation focuses on multiple-choice reasoning, and future work should extend confidence–rationale alignment to open-ended and retrieval-intensive reasoning tasks.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

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

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*.

Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan. Association for Computing Machinery.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan

Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Mistral AI. 2025. Ministral 3 8b reasoning 2512. Hugging Face Model Card.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

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

Qwen Team. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ali Razghandi, Seyed Mohammad Hadi Hosseini, and Mahdieh Soleymani Baghshah. 2025. CER: Confidence enhanced reasoning in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7918–7938, Vienna, Austria. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. 2024. Thermometer: Towards universal calibration for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 44687–44711. PMLR.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence*, 7:221–231.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust LLMs: Aligning confidence with response quality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5984–5996, Bangkok, Thailand. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasović, and Yonatan Belinkov. 2025. Measuring chain of thought faithfulness by unlearning reasoning steps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9935–9960, Suzhou, China. Association for Computational Linguistics.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18138, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488.

Boxuan Zhang and Ruqi Zhang. 2025. CoT-UQ: Improving response-wise uncertainty quantification in LLMs with chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26114–26133, Vienna, Austria. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

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

A Additional Implementation Details

A.1 Committed-Answer Confidence Extraction

For each generated completion, we first parse the committed answer y from the final answer line `Final Answer: X`. We then compute confidence over the discrete answer-option labels rather than using the model's verbalized confidence. Specifically, after the model generates a rationale and a final-answer line, we locate the generated answer label in the token sequence and take the exact generated prefix before that label as the scoring context. Under this prefix, we score each candidate answer label $\ell \in \mathcal{Y}$ as a continuation.

In our implementation, candidate labels are scored with a leading-space template `X`, where X is one of the valid option labels. Thus, for a four-choice dataset, the candidate continuations are `A`, `B`, `C`, and `D`. For MedQA and MathQA, we additionally include `E`. Let s_ℓ denote the conditional log-probability of candidate continuation ℓ under the exact generated prefix. We convert these scores into an option-level probability distribution:

$$p(\ell \mid x, r) = \frac{\exp(s_\ell)}{\sum_{\ell' \in \mathcal{Y}} \exp(s_{\ell'})}.$$

The committed-answer confidence is then defined as

$$C = p(y \mid x, r),$$

where y is the parsed answer selected by the model. Since the candidate continuations are fixed answer labels rather than full option texts, they are short and comparable across options; we therefore use the summed continuation log-probability without additional length normalization.

No temperature scaling is applied when computing these option-level probabilities. Although generation may use sampling parameters during response generation, the confidence score is computed post hoc from the model's raw conditional likelihoods under the generated prefix. The resulting distribution is stored as `choice_probs`, and the probability assigned to the parsed answer is stored as `p_parsed`. This `p_parsed` value is used as $C \in [0, 1]$ for ECE, Brier score, the confidence–rationale alignment reward, and the alignment error

metric. If a completion does not contain a parseable final answer, it is counted as incorrect and assigned zero rationale-support score. This procedure measures the model's probability for the committed answer label under the generated reasoning context, rather than the probability of the full option text or any self-reported confidence statement.

A.2 Reproducibility Details

For reward computation during training, we use GPT-OSS 20B as the sole rationale-support judge. The judge input excludes gold option labels, official solutions, official explanations, and gold rationales; the auxiliary background knowledge is used only to assess factual grounding. The general support score Q_{gen} is computed from the rubric axes in Table 1. The format-validity axis G_1 is used as a hard gate. For the remaining general axes, we use weights of 0.20, 0.30, 0.20, 0.25, and 0.05 for task understanding, evidence grounding, inference coherence, answer bridge, and structure, respectively. The task-specific score Q_{task} is computed by averaging the two task-specific axes. We set $w_{\text{gen}} = 0.75$ and $w_{\text{task}} = 0.25$, apply the penalty P for severe structural failure or unsupported reasoning, and clip the final score to $[0, 1]$. The training process uses 3 NVIDIA H100 GPUs.

For the confidence–rationale alignment score in Equation 3, we set $\alpha = 0.6$, $\beta_{\text{over}} = 0.6$, and $\beta_{\text{under}} = 0.4$, assigning a larger penalty to confidence that exceeds rationale support. For the final reward in Equation 4, we use $\lambda = 0.2$, $\gamma = 0.2$, and $\beta = 0.1$. We train with LoRA using rank 16, LoRA alpha 32, and dropout 0.05. Each model is trained for one epoch with learning rate 10^{-5}, batch size 8, group size 8 for GRPO sampling, and KL coefficient 0.04.

A.3 Prompt Templates

Generation Prompt. The policy model is prompted with the following template:

```
You are a careful reasoner.

Answer the following {dataset_display}
multiple-choice question. Give a short
chain-of-thought in several steps.
Then output Final Answer as a single
letter from {choice_list}. Stop immediately
after the Final Answer line.

{fact_block}Question: {question}
Options:
{options_text}

Use exactly this format:
Chain-of-Thought:
1) ...
2) ...
Final Answer: {choice_template}
```

Table 5: Linguistic features used for correctness prediction. Features are extracted from the generated CoT reasoning trace.

Feature	Description
Length	Number of characters and words in the generated trace.
Step count	Number of explicit numbered reasoning steps.
Sentence length	Average number of words per sentence.
Hedging	Density of uncertainty markers, e.g., "maybe", "likely".
Connectives	Density of causal/logical connectives, e.g., "therefore".
Backtracking	Density of revision markers, e.g., "wait", "actually".
Digit density	Fraction of characters that are digits.
Repetition	Duplicate trigram rate in the generated trace.
Lexical diversity	Ratio of unique words to total words.

Judge System Prompt. The rubric judge is prompted with the following system instruction:

```
Reasoning: low

You are an evaluator for multiple-choice
question reasoning. You are given:
(a) a dataset name
(b) a question and answer options
(c) optional reference/evidence
(d) a CHAIN-OF-THOUGHT answer written
    by another model

Score the CHAIN along the 8 rubric axes.

Use reliable domain knowledge when needed:
- OpenBookQA: science/common-sense knowledge
- MedQA: medical knowledge
- MathQA: mathematical reasoning/computation

You are NOT given the correct answer.
Do not use a gold label. Judge only from
the question, options, optional evidence,
and chain.

For each axis, output the value only.
No evidence quotes. Output STRICT JSON only.
```

Judge Input Prompt. For each generated reasoning trace, the judge receives:

```
DATASET: {dataset_name}

QUESTION: {question}

OPTIONS:
A: {option_A}
B: {option_B}
...

PROVIDED REFERENCE/EVIDENCE:
{fact1}

CHAIN-OF-THOUGHT:
{chain}

CHOSEN ANSWER: {parsed_answer}

Evaluate this chain along the 8 rubric axes.
Output STRICT JSON only.
```

The reference-evidence block is omitted when no reference evidence is used.

B Linguistic Features for Correctness Prediction

C Additional Judge Evaluation

Table 6: **Rationale-support quality and confidence–rationale alignment results evaluated by GPT-OSS 20B**. Q measures the quality of the generated reasoning trace. E_{align} denotes the average confidence–rationale alignment error. ↑ indicates higher is better, and ↓ indicates lower is better. All values are reported as percentages.

		Ministral 3 8B		**Qwen2.5 7B**		**Gemma 2 9B**	
Dataset	**Method**	$Q \uparrow$	$E_{\text{align}} \downarrow$	$Q \uparrow$	$E_{\text{align}} \downarrow$	$Q \uparrow$	$E_{\text{align}} \downarrow$
MathQA	Base	94.51	6.72	83.67	13.61	75.01	21.11
	SFT	94.77	5.42	82.41	14.29	76.75	21.20
	GRPO	96.19	5.14	84.93	11.50	76.59	20.06
	CoRA (Ours)	**97.09**	**2.97**	**87.70**	**9.54**	**79.07**	**16.75**
MedQA	Base	81.31	18.15	59.30	36.01	69.83	29.28
	SFT	81.92	17.48	61.26	34.79	71.54	28.48
	GRPO	**83.89**	**16.28**	**64.89**	**28.56**	72.01	26.84
	CoRA (Ours)	82.17	16.87	63.78	32.29	**78.72**	**20.42**
OpenBookQA	Base	82.17	5.98	90.89	9.97	92.02	7.25
	SFT	95.08	4.97	91.38	8.43	91.90	7.77
	GRPO	94.81	5.51	90.95	9.43	93.10	6.59
	CoRA (Ours)	**95.48**	**4.24**	**92.41**	**7.73**	**93.25**	**6.15**

Table 7: **Rationale-support quality and confidence–rationale alignment results evaluated by Qwen 3 32B.** Q measures the quality of the generated reasoning trace, while E_{align} denotes the average confidence–rationale alignment error. ↑ indicates higher is better, and ↓ indicates lower is better. All values are reported as percentages.

		Ministral 3 8B		**Qwen2.5 7B**		**Gemma 2 9B**	
Dataset	**Method**	$Q \uparrow$	$E_{\text{align}} \downarrow$	$Q \uparrow$	$E_{\text{align}} \downarrow$	$Q \uparrow$	$E_{\text{align}} \downarrow$
MathQA	Base	96.51	5.10	84.99	12.49	78.83	18.54
	SFT	95.50	4.62	84.50	12.77	80.24	17.97
	GRPO	96.42	5.32	86.59	10.43	79.26	18.10
	CoRA (Ours)	**98.68**	**1.31**	**88.33**	**9.49**	**82.98**	**14.29**
MedQA	Base	88.84	11.41	67.23	30.98	72.12	27.29
	SFT	89.51	10.39	70.57	27.83	74.08	25.86
	GRPO	**90.99**	**9.35**	71.12	**24.08**	74.63	24.57
	CoRA (Ours)	90.46	10.13	**72.36**	25.97	**84.09**	**15.70**
OpenBookQA	Base	93.37	7.44	91.12	10.14	89.68	9.94
	SFT	**95.22**	**4.72**	92.24	8.22	90.73	9.36
	GRPO	95.04	5.59	91.36	9.16	91.76	8.14
	CoRA (Ours)	94.38	5.30	**92.25**	**7.68**	**92.96**	**7.08**