CALIBRATING VERBALIZED CONFIDENCE WITH RUBRICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can assess the correctness of their own responses through verbalized confidence, helping users decide whether to adopt the response or seek further review. However, models often exhibit overconfidence when their responses are incorrect, limiting the practical value of these estimates. Existing research addresses this problem by refining how confidence judgments are elicited and aggregated, yet whether expressed confidence is reliable under more stringent and explicit verification criteria remains underexplored. In this paper, we systematically analyze models' verbalized confidence through progressively stricter rubric-based checks and propose two confidence calibration methods. The proposed methods integrate verification signals from these checks with the model's holistic self-assessment, thereby mitigating overconfidence. Experiments on eight benchmarks spanning diverse reasoning tasks and domain knowledge, across three model families and three model scales ranging from 8B to 32B, demonstrate improved calibration while preserving discrimination between correct and incorrect responses. Further analysis shows that rubric-based verification provides useful signals beyond direct confidence, and that the gains cannot be attributed merely to more conservative confidence estimates.

1 INTRODUCTION

Large language models (LLMs) can not only generate responses but also assess the correctness of their own responses (Kadavath et al., 2022). Such self-assessment is often expressed as a confidence score that guides decisions about whether to accept or further verify a response. Therefore, the value of such scores for decision-making depends on how well they reflect the actual likelihood of correctness (Guo et al., 2017). Among existing confidence estimation methods, *verbalized confidence* is elicited by directly prompting the model to report a numerical estimate of correctness (Tian et al., 2023; Xiong et al., 2024). This approach requires no access to model internals, making it practical for deployed LLMs in black-box settings. In practice, however, these estimates are often overconfident, limiting their reliability (Wang & Stengel-Eskin, 2026).

Recent work has improved calibration by varying prompting and sampling conditions and aggregating the resulting confidence judgments (Zhou et al., 2025; Wang & Stengel-Eskin, 2026; Geng et al., 2024; Shorinwa et al., 2025). In particular, variation and agreement across multiple judgments provide useful signals for refining confidence estimates and highlight that verbalized confidence depends not only on the response itself but also on how it is evaluated. This has motivated assessments under semantically equivalent prompts and answer variations (Xia et al., 2026). However, this line of work mainly focuses on how confidence judgments vary or remain consistent across evaluation conditions, without explicitly examining whether the model's original confidence remains justified as verification becomes progressively more stringent. This motivates complementing repeated confidence judgments with explicit checks of the reasoning process and final conclusion.

To this end, we introduce rubrics (Hashemi et al., 2024; Siro et al., 2026) into confidence estimation and evaluate the response across a sequence of verification levels ranging from basic plausibility to full verification. At each level, the model estimates the probability that the response satisfies the corresponding requirement. The resulting *rubric profile* captures how the model's assessment changes as the verification requirements become more demanding. Specifically, lower scores at stricter levels can expose uncertainty that is obscured by the original confidence estimate. Our

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

preliminary analysis with Qwen3-8B illustrates this behavior (Section 3.1). Incorrect responses can receive consistently high direct confidence across prompt paraphrases, whereas the average gap in rubric scores between correct and incorrect responses widens across verification levels.

Building on these observations, we obtain multiple direct confidence estimates and rubric scores through prompt paraphrases while keeping the response and verification requirements fixed. The mean direct confidence summarizes the model's holistic judgment of response correctness, whereas the mean rubric score summarizes the model's assessments against the verification criteria. We combine these signals in *Rubric Calibration* (RC), using rubric support as a response-specific correction factor: weaker verification support results in a greater reduction in direct confidence. We further observe that substantial overconfidence can persist when relative rubric profiles are highly repeatable across paraphrases or when the RC confidence remains high. Accordingly, we introduce *Adaptive Rubric Calibration* (RC-A), which considers both rubric profile repeatability and RC confidence and applies an additional rubric correction when either condition is met. RC and RC-A reuse the same direct confidence estimates and rubric scores, differing only in how these signals are combined.

We evaluate the proposed methods on eight benchmarks spanning diverse reasoning tasks and domain knowledge, using three model families. The results show that our methods improve calibration on average compared with a broad set of baselines while maintaining competitive discrimination. We further use controlled comparisons to isolate the contribution of rubric information, showing that the observed gains are not explained solely by more conservative confidence estimates. Overall, rubric assessments provide effective response-specific signals that complement holistic confidence judgments in calibration.

2 Related Work

2.1 Confidence Estimation and Calibration

LLM confidence can be estimated from either internal model signals or generated outputs. Internal signals include token probabilities, hidden representations, and attention scores (Kadavath et al., 2022). Generation-based methods instead infer confidence from text generated by the model, without requiring access to these internal signals. Verbalized confidence is one such approach, in which the model expresses uncertainty through a numerical score. Prior work shows that models can be trained to report calibrated confidence (Lin et al., 2022), while other approaches improve calibration through inference-time elicitation and post-hoc score adjustment (Tian et al., 2023; Zhou et al., 2025; Wang & Stengel-Eskin, 2026). Complementary work examines how information from the model's reasoning and responses can inform confidence estimation. Some methods incorporate self-reflective rationales and contrasts between answers into training, linking confidence estimates more closely to reasoning and response content (Xu et al., 2024; Seo et al., 2026). Others combine multiple confidence judgments elicited under different evaluation conditions with agreement across sampled answers, aggregating or normalizing these signals to reduce reliance on any single judgment (Xiong et al., 2024).

2.2 Rubric-Based LLM Evaluation

Rubric-based evaluation structures LLM judgments through explicit assessment criteria for each response. Prior work uses criterion-guided scoring and multidimensional assessments to improve agreement between automated evaluation and human judgments (Liu et al., 2023; Hashemi et al., 2024). More recent work focuses on constructing and refining these criteria, including their decomposition and filtering to improve coverage and reduce redundancy (Siro et al., 2026; Shen et al., 2026). Other studies examine whether generated criteria reflect human standards (Zhou et al., 2026), whether their application is grounded in verifiable evidence (Hong et al., 2026), and whether structured feedback can serve as a reward signal for reinforcement learning (Gunjal et al., 2026). Overall, by making the requirements for a satisfactory response explicit, rubrics provide a structured basis for analyzing the support behind a model judgment. Accordingly, we introduce rubrics into model confidence estimation and use rubric scores to complement the model's holistic judgment when calibrating verbalized confidence.

3 RUBRIC-BASED CONFIDENCE CALIBRATION

Problem setting. Given a problem x, let a denote an LLM-generated candidate solution. The goal is to estimate a calibrated confidence score $C(x, a) \in [0, 1]$ for the correctness of a by querying the same LLM. For a controlled comparison across confidence estimation methods, the problem x and candidate solution a are held fixed. Correctness labels are used only to evaluate the resulting confidence estimates, not for confidence elicitation or score computation.

3.1 MOTIVATING OBSERVATIONS

The analysis covers 3,910 responses generated by Qwen3-8B across all eight benchmarks. For each response, the same model provides direct confidence estimates and rubric scores using four prompt paraphrases for each elicitation type. The rubric spans five verification levels, from surface validity to full verification (Section 3.2). Scores are averaged across paraphrases for each response, and the analysis pools responses across all benchmarks. The corresponding analysis for Gemma-3-12B is provided in Appendix D.1.

Repeated judgments can remain overconfident. Figure 1(a) shows that direct confidence is concentrated near the upper end of the scale for both correct and incorrect responses. The mean is 0.851 for incorrect responses, compared with 0.969 for correct ones. Overconfidence persists even when the judgments agree exactly: 1,208 responses receive a score of 0.950 under all four paraphrases, yet only 57.7% of them are correct. Thus, agreement among repeated holistic judgments does not by itself ensure calibration.

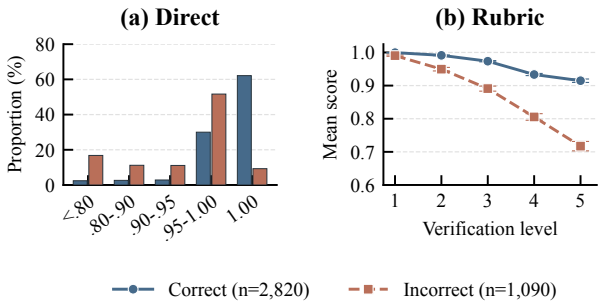


Figure 1: Qwen3-8B on eight benchmarks. (a) Within-group proportions by confidence category. (b) Mean rubric scores, with 95% bootstrap confidence intervals; levels 1–5 correspond to the verification levels defined in Section 3.2.

Stricter verification increases separation between correct and incorrect responses. Figure 1(b) shows that correct and incorrect responses both receive near-maximal scores for surface validity: 0.999 and 0.990, respectively. At full verification, the corresponding means fall to 0.915 and 0.717. The gap widens across verification levels, from 0.009 at the first level to 0.198 at the last. This widening separation motivates using rubric scores as a complement to direct confidence. Section 5.1 further analyzes the response-specific information provided by rubric scores through controlled comparisons, distinguishing its contribution from uniform confidence reduction.

3.2 RUBRIC-BASED CONFIDENCE ELICITATION

Explicit evaluation criteria can structure LLM judgments beyond a holistic assessment (Liu et al., 2023; Hashemi et al., 2024), while stepwise self-verification can expose errors within a reasoning chain (Miao et al., 2024). In this work, we organize the verification requirements into a rubric with K progressively stringent levels, moving from surface validity to full verification of the generated response. Our implementation uses five levels ($K = 5$), described below. The rubric begins with *surface validity*, which considers relevance, the plausibility of the answer format, and the absence of obvious flaws. *Local correctness* further examines calculations, factual claims, and individual inference steps, while *logical sufficiency* assesses whether the stated reasoning supports the conclusion without critical gaps or unjustified leaps. *Adversarial robustness* extends the examination to hidden constraints, edge cases, and potential counterexamples. Finally, *full verification* requires the entire solution to be correct and the final answer to follow from the problem under the strictest standard. At each level, the score represents the model's estimate of the probability that the candidate solution satisfies the corresponding requirement. The verification levels are ordered from less to more stringent. We retain the model's original scores even when they do not decrease monotonically across levels.

For each elicitation type, we use V same-language prompt paraphrases ($V = 4$ in our main experiments) while holding the problem, candidate solution, and rubric criteria fixed. For paraphrase index v, direct elicitation returns a single correctness probability $d_v \in [0, 1]$, whereas rubric elicitation returns $\mathbf{r}_v = (r_{v1}, \ldots, r_{vK}) \in [0, 1]^K$. The direct and rubric prompts are paraphrased independently, so their indices do not indicate a one-to-one correspondence. The original direct confidence and rubric scoring prompts are provided in Appendix A.5.

3.3 Our Methods

Rubric Calibration (RC). Aggregating repeated assessments of the same input can reduce reliance on a single elicitation and is a common strategy in uncertainty estimation (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Gawlikowski et al., 2023). For rubric assessment, averaging across verification levels summarizes the full profile without relying on a single level. Accordingly, for each fixed problem–solution pair (x, a), we separately aggregate the multiple prompt paraphrases used for direct confidence and rubric assessment:

$$D_{\text{mean}} = \frac{1}{V} \sum_{v=1}^{V} d_v, \qquad R_{\text{mean}} = \frac{1}{VK} \sum_{v=1}^{V} \sum_{k=1}^{K} r_{vk}. \tag{1}$$

These means capture complementary aspects of the candidate solution: D_{mean} summarizes the model's holistic judgment of correctness, whereas R_{mean} aggregates the support reported under the verification requirements. Rather than treating R_{mean} as a second confidence estimate to be averaged with D_{mean}, we propose RC, which uses R_{mean} as a response-specific multiplicative correction factor for verbalized confidence (Zhou et al., 2025):

$$C_{\text{RC}}(x, a) = D_{\text{mean}}(x, a)\, R_{\text{mean}}(x, a). \tag{2}$$

The resulting RC confidence largely preserves the model's holistic judgment when rubric support is strong and decreases as that support weakens. Consequently, solutions with similar direct confidence can receive different RC confidence values.

Rubric profile repeatability. The rubric provides a structured profile, allowing comparison of relative support across verification levels rather than only overall scores. To quantify how consistently these relative patterns recur across paraphrases, the profile from each paraphrase is centered as $\tilde{\mathbf{r}}_v = \mathbf{r}_v - \mu_v \mathbf{1}$, where $\mu_v = K^{-1} \sum_{k=1}^{K} r_{vk}$. Let q_{uv} be the cosine similarity between $\tilde{\mathbf{r}}_u$ and $\tilde{\mathbf{r}}_v$, with a value of zero when either profile has zero centered norm. The resulting coefficient is

$$A_R = \max\left(0, \frac{1}{\binom{V}{2}} \sum_{u<v} q_{uv}\right), \qquad A_R \in [0, 1]. \tag{3}$$

Residual overconfidence. Although RC uses rubric support to correct direct confidence, RC confidence can remain substantially higher than empirical accuracy within some subsets of responses. We further examine how this residual overconfidence relates to rubric profile repeatability. For example, using 0.9 as the profile repeatability threshold, Qwen3-8B responses with $A_R \geq 0.9$ have a mean confidence of 0.804 but an accuracy of only 0.529. Although this group comprises 54.42% of all responses, 92.02% of all incorrect responses fall within it. Figure 2 shows both the concentration of errors and the remaining confidence–accuracy gap for Qwen3-8B and Gemma-3-12B. Corresponding statistics for Llama-3.1-8B are given in Appendix D.2. These results show that high A_R can identify responses that remain substantially overconfident after RC and can therefore trigger further correction.

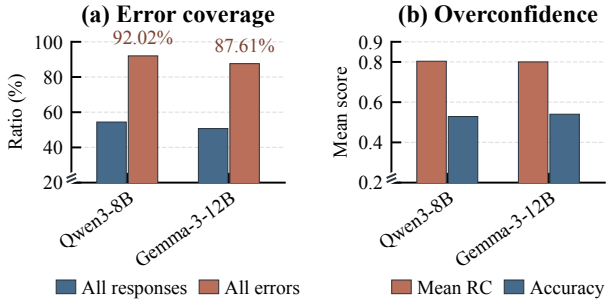


Figure 2: Residual overconfidence among responses with $A_R \geq 0.9$ across eight benchmarks. (a) Response and error shares, with labels showing error coverage. (b) Mean RC confidence and accuracy.

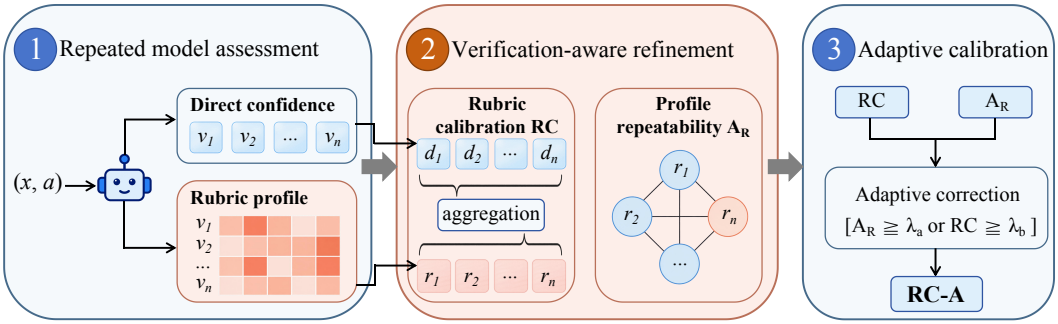


Figure 3: Overview of the proposed rubric-based confidence calibration method. Multiple prompt paraphrases elicit direct confidence estimates and rubric profiles. Rubric Calibration (RC) combines the two signals, while Adaptive Rubric Calibration (RC-A) applies an additional rubric correction based on rubric profile repeatability and RC confidence.

Profile repeatability alone, however, does not cover all residual overconfidence. Taking 0.9 as an illustrative threshold for C_{RC} as well, Qwen3-8B responses with $A_R < 0.9$ and $C_{\mathrm{RC}} \geq 0.9$ have a mean confidence of 0.994, exceeding their accuracy of 0.954. Thus, even below the repeatability threshold, responses with high RC confidence may benefit from further correction. RC-A therefore treats high profile repeatability and high RC confidence as complementary conditions, applying one additional rubric correction when either condition holds. Detailed group statistics are provided in Appendix D.2.

Adaptive Rubric Calibration (RC-A). Based on these observations, RC-A applies one additional rubric correction when either the profile repeatability or the RC confidence is high:

$$g(x,a) = \mathbf{1}[A_R(x,a) \geq \lambda_a \ \vee \ C_{\mathrm{RC}}(x,a) \geq \lambda_b]\,,$$
$$C_{\text{RC-A}}(x,a) = C_{\mathrm{RC}}(x,a)\, R_{\mathrm{mean}}(x,a)^{\,g(x,a)}. \tag{4}$$

Here, the additional rubric correction is applied only once, even when both conditions hold. The thresholds λ_a and λ_b determine which responses receive the additional correction; their experimental settings are specified in Section 4.1. RC and RC-A use the same model outputs and therefore require the same number of model calls. We further examine the contribution of rubric information itself and the effect of the adaptive gate in Section 5.1 and the ablation study in Section 4.3, respectively. Figure 3 summarizes the proposed method.

4 Experiments

4.1 Experimental Setup

Benchmarks and models. The evaluation spans eight benchmarks covering diverse reasoning demands and knowledge requirements, following the task coverage considered in Xiong et al. (2024). GSM8K (Cobbe et al., 2021) and AIME25 (MathArena, 2025) examine mathematical reasoning, ranging from grade-school word problems to competition mathematics. Beyond mathematics, Date Understanding and Object Counting from BIG-Bench Hard (Suzgun et al., 2023) assess symbolic reasoning, while StrategyQA (Geva et al., 2021) and GPQA Diamond (Rein et al., 2023) require commonsense and scientific reasoning, respectively. The Business Ethics and Professional Law subsets of MMLU (Hendrycks et al., 2021) further extend the evaluation to ethical judgment and professional knowledge. These benchmarks comprise 3,910 questions in total. The main comparison covers three model families through Qwen3-8B (Yang et al., 2025), Gemma-3-12B-IT (Gemma Team et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). The model size analysis additionally includes Qwen3-14B and Qwen3-32B, allowing comparison across three model sizes within the Qwen3 family. Sample counts and implementation details are provided in Appendix A.

Baselines. The comparison includes eight confidence estimators, adapted where necessary to evaluate the same fixed candidate solution. Direct (Tian et al., 2023) elicits confidence without an explanation, while CoT (Kojima et al., 2022; Xiong et al., 2024) allows reasoning before the confidence

Table 1: Confidence estimation across eight benchmarks (macro averages; all metrics are scaled by 10^2). **Bold** and underlined values indicate the best and second-best results, respectively, for each model and metric.

Method	Qwen3-8B			Gemma-3-12B			Llama-3.1-8B		
	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑
Direct	21.80	21.36	69.33	25.98	24.85	62.10	35.61	35.19	62.67
CoT	22.57	21.86	70.01	24.10	22.93	62.98	25.57	28.15	63.06
Multi-Step	22.32	21.16	68.29	28.79	26.71	58.41	35.10	34.80	60.50
Self-Probing	16.89	18.28	68.37	17.48	18.82	65.31	22.32	26.24	63.19
Sampling	22.27	21.58	70.85	25.97	24.61	63.95	35.07	34.22	65.89
Paraphrase Mean	21.93	21.14	72.82	24.96	23.63	63.87	33.88	33.35	66.65
Multilingual Mean	20.65	20.07	75.11	26.42	25.07	65.50	36.02	35.16	63.34
SteerConf	15.29	16.49	**75.76**	19.05	19.17	**67.34**	23.83	25.42	69.54
RC (Ours)	16.57	17.17	75.67	17.54	18.00	65.99	18.66	21.57	69.49
RC-A (Ours)	**12.91**	**15.04**	75.45	**13.12**	**15.58**	66.24	**18.19**	**20.37**	**69.97**

estimate. Multi-Step, adapted from Xiong et al. (2024), requests step-level assessments and uses the final confidence. Self-Probing and Sampling (Xiong et al., 2024) are selected to represent error-based probing and repeated stochastic evaluation, respectively. Paraphrase Mean and Multilingual Mean serve as elicitation controls, averaging scores obtained under same-language prompt paraphrases and translations into different languages, respectively. SteerConf (Zhou et al., 2025) combines confidence estimates elicited under five steering conditions. The baseline implementations are described in Appendix A.3. RC and RC-A use a five-level rubric and four English prompt paraphrases for each elicitation type. RC-A uses a shared threshold of 0.9 for both conditions across all models and benchmarks, without model- or task-specific values. Appendices C.1 and C.2 examine threshold sensitivity and the number of paraphrases, respectively.

Evaluation metrics. Confidence estimates are evaluated from two complementary perspectives: calibration and error discrimination (Xiong et al., 2024; Wang & Stengel-Eskin, 2026). For calibration, expected calibration error (ECE) (Naeini et al., 2015) measures agreement between confidence and accuracy using ten equal-width bins, while the Brier score (Brier, 1950) measures squared error against the correctness labels. For error discrimination, AUROC (Fawcett, 2006) assesses how well confidence distinguishes correct from incorrect responses. The main comparison uses macro averages, giving each benchmark equal weight. PR-P and PR-N are reported in Appendix B; metric definitions are given in Appendix A.4.

4.2 Main Results

Table 1 summarizes macro averages across the eight benchmarks for RC, RC-A, and the eight baselines. RC-A achieves lower ECE and Brier scores than every baseline on all three model families, including methods that aggregate multiple judgments. Relative to the strongest baseline for each model and metric, its absolute reductions range from 2.37 to 4.36 for ECE and from 1.45 to 5.05 for Brier. The simpler RC estimator also improves both metrics over Direct and Paraphrase Mean on all three models, showing that rubric-based correction is useful even without the adaptive gate.

On Llama-3.1-8B, RC-A further reduces ECE from 18.66 to 18.19 and Brier from 21.57 to 20.37 relative to RC. Both variants also provide competitive discrimination between correct and incorrect responses. RC attains an AUROC of 75.67 on Qwen3-8B, close to SteerConf's 75.76; RC-A ranks second on Gemma-3-12B and achieves the highest AUROC on Llama-3.1-8B at 69.97. The extra correction does not uniformly improve discrimination: on Qwen3-8B, RC-A's AUROC of 75.45 is slightly below RC's 75.67. Its primary benefit is calibration, while discrimination remains competitive.

Table 2: Component ablations across eight benchmarks (macro averages; all metrics are scaled by 10^2). A_R only and C_{RC} only apply the extra correction R_{mean} when the corresponding quantity is at least 0.9. **Bold** and underlined values indicate the best and second-best results, respectively, for each model and metric.

Variant	Qwen3-8B			Gemma-3-12B			Llama-3.1-8B		
	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑
Direct Mean	22.14	21.37	71.71	25.06	23.72	63.85	33.42	32.98	66.71
Rubric Mean	22.56	21.20	75.21	21.75	20.65	65.80	21.86	23.40	68.89
RC	16.57	17.17	**75.67**	17.54	18.00	65.99	18.66	21.57	69.49
A_R only	13.10	15.11	75.32	13.40	15.67	66.22	18.53	20.40	69.79
C_{RC} only	15.93	16.96	74.94	17.04	17.83	65.13	18.33	21.54	69.66
RC-A	**12.91**	**15.04**	75.45	**13.12**	**15.58**	**66.24**	**18.19**	**20.37**	**69.97**

4.3 Ablation Study

Table 2 separates the contributions of the two score branches and the two gate conditions, using the same candidate solutions and elicited scores throughout. Direct Mean and Rubric Mean retain only D_{mean} and R_{mean}, respectively. Both variants use the same four prompt paraphrases as the proposed methods. Rubric Mean has higher AUROC than Direct Mean on all three models. Combining them in RC further improves ECE, Brier, and AUROC over either branch alone, supporting the use of verification assessments alongside direct confidence. For Qwen3-8B, for example, AUROC increases from 71.71 for Direct Mean and 75.21 for Rubric Mean to 75.67 for RC.

To isolate the gate conditions, A_R only applies the additional rubric correction when $A_R \geq 0.9$, whereas C_{RC} only applies it when $C_{\mathrm{RC}} \geq 0.9$. Thus, each removes one condition from RC-A while retaining the other at the same threshold. The adaptive gate yields lower ECE and Brier and higher AUROC than either single-condition variant on all three models. Adding the high-confidence condition to A_R only reduces ECE from 13.10 to 12.91 on Qwen3-8B, from 13.40 to 13.12 on Gemma-3-12B, and from 18.53 to 18.19 on Llama-3.1-8B. Relative to RC, C_{RC} only yields a smaller Brier reduction than A_R only on all three models. The two conditions therefore provide complementary corrections in these aggregate results, although the incremental Brier and AUROC gains over A_R only are small.

5 Analysis and Extensions

5.1 The Value of Rubric Scores

Previous experimental results show that combining direct confidence with rubric scores improves confidence estimation. However, since the rubric factor is at most one, this combination also lowers confidence. It is therefore important to distinguish the benefit of confidence reduction from the information that rubric scores provide about individual responses. Figure 4 examines these two effects on Qwen3-8B, using RC without the additional adaptive gate. The four columns compare Direct Mean, two controls, and RC; the upper row shows confidence distributions separated by correctness, and the lower row shows reliability diagrams. Complete comparisons across all five metrics for all three models appear in Appendix D.3.

Matched Scaling replaces each response's rubric factor with a model-specific constant, giving $C_{\mathrm{scale},i} = \gamma D_{\mathrm{mean},i}$. A single γ is shared across all benchmarks of a model and chosen to match RC's mean confidence averaged equally across datasets, without using correctness labels. Under this control, RC achieves a lower Brier score and a higher AUROC on all three models. Thus, uniformly lowering scores to the same level does not reproduce the full benefit of rubric-based correction.

Shuffled Pairing further controls for the confidence distribution by randomly exchanging rubric profiles among responses from the same benchmark with exactly equal direct confidence means. The resulting estimate is $C_{\mathrm{shuffle},i} = D_{\mathrm{mean},i} R_{\mathrm{mean},\pi(i)}$, where $D_{\mathrm{mean},i} = D_{\mathrm{mean},\pi(i)}$. Each response's entire set of rubric scores moves together, while singleton groups remain unchanged; neither the grouping nor the exchange uses correctness labels. This preserves every confidence value and its

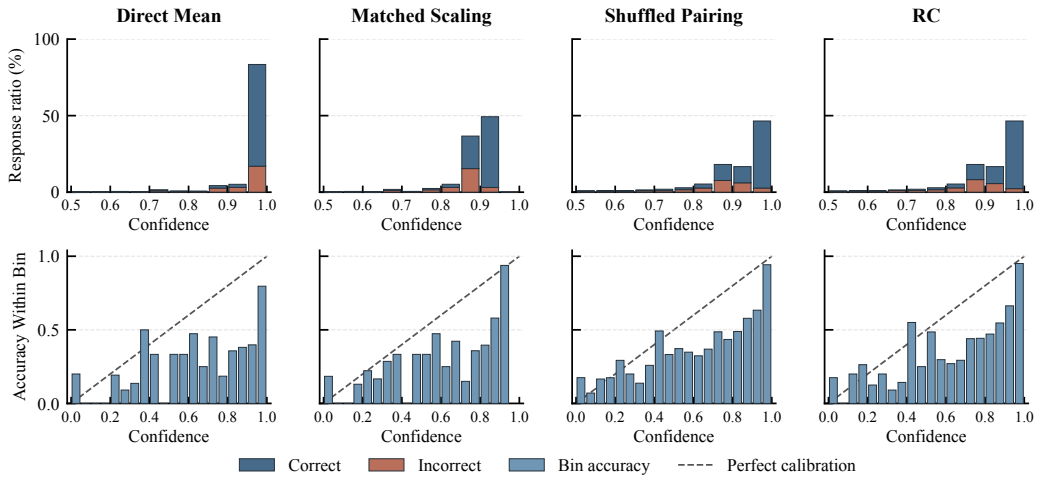


Figure 4: Confidence distributions (top) and reliability diagrams (bottom) for Qwen3-8B across eight benchmarks. Stacked bars show correct and incorrect responses as proportions of all responses. Reliability diagrams use bin centers over the full confidence range, with the dashed diagonal indicating perfect calibration.

frequency within each benchmark while changing only the rubric assignment to each response. Accordingly, the Shuffled Pairing and RC columns have identical total histogram heights, but different proportions of correct and incorrect responses within the bins. For Qwen3-8B, there are 2,471 responses with confidence in $[0.9, 1]$ under either assignment, but their accuracy is 87.45% under RC versus 86.06% on average after shuffling. Across 2,000 random permutations, mean macro AUROC decreases from 75.67 to 71.51 on Qwen3-8B, from 65.99 to 64.05 on Gemma-3-12B, and from 69.49 to 65.93 on Llama-3.1-8B. Brier also increases on all three models. Even when mean direct confidence is fixed, rubric scores therefore help distinguish correct from incorrect responses.

These comparisons support using rubric scores as a response-specific correction rather than merely a means of lowering confidence uniformly. The evidence is more consistent for Brier and discrimination than for ECE: RC has lower macro ECE than both controls on Qwen3-8B and Gemma-3-12B, but not on Llama-3.1-8B. The original pairing thus provides useful information without necessarily improving every calibration metric.

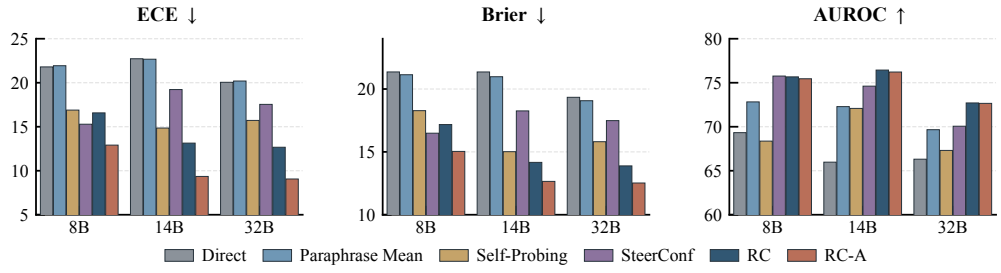


Figure 5: Confidence estimation across Qwen3 model sizes (macro averages; all metrics are scaled by 10^2). Results are shown for Qwen3-8B, Qwen3-14B, and Qwen3-32B. Selected baselines are shown for readability; complete results are reported in Table 7.

5.2 Scalability Across Model Sizes

To examine whether the benefits extend to larger models, the evaluation includes Qwen3-14B and Qwen3-32B alongside Qwen3-8B. The rubric, four English prompt paraphrases, and aggregation rules remain unchanged, with $\lambda_a = \lambda_b = 0.9$ at every size. Each model assesses its own candidate solutions, following the main experimental protocol. Figure 5 summarizes the macro averages across the same eight benchmarks; Appendix B.3 reports all eight baselines.

RC-A achieves the lowest ECE and Brier score at all three sizes. For Qwen3-32B, it reduces ECE from 15.72 to 9.06 and Brier from 15.81 to 12.52 relative to Self-Probing, the strongest baseline on these two metrics. The simpler RC estimator also outperforms all baselines in ECE, Brier, and AUROC at both 14B and 32B, while RC-A further improves its calibration. Their AUROC remains competitive: at 32B, RC and RC-A attain 72.70 and 72.65, respectively, compared with 72.46 for the strongest baseline, Sampling. This shared configuration improves macro-average calibration across the evaluated model families and sizes, demonstrating the applicability of the same correction rule across different settings.

5.3 Rubric Calibration Beyond Direct Confidence

To examine whether rubric information benefits other confidence estimators, we apply rubric calibration to CoT and SteerConf, which use explicit reasoning and confidence steering, respectively. Their prompting procedures and confidence estimates remain unchanged. Rubric scores are elicited independently using the main experiments' four English prompt paraphrases. For a base confidence B, RC gives $C_{\mathrm{RC},B} = BR_{\mathrm{mean}}$. RC-A applies an additional rubric factor when $A_R \geq 0.9$ or $C_{\mathrm{RC},B} \geq 0.9$. Each base estimator and its extensions use identical candidate solutions, with the RC-A thresholds unchanged.

Table 3 shows that both extensions improve ECE, Brier, and AUROC over their respective base estimators across all three models. On Qwen3-8B, CoT with RC-A reduces ECE from 22.57 to 14.03 and increases AUROC from 70.01 to 73.62. The gains also extend to SteerConf: on Gemma-3-12B, RC-A reduces its ECE from 19.05 to 9.80 and Brier from 19.17 to 14.83, while maintaining a slightly higher AUROC. The additional adaptive correction is not uniformly preferable to RC; on Llama-3.1-8B, RC yields lower ECE and Brier for both base estimators. These findings extend the benefits of rubric calibration to confidence elicited through reasoning or steering. Each extension requires four additional rubric calls per candidate solution, making this a comparison of applicability rather than equal-cost performance. Appendix B.4 reports complete metrics, and Appendix C examines parameter sensitivity and inference budget.

Table 3: Rubric calibration applied to CoT and SteerConf across eight benchmarks (macro averages; all metrics are scaled by 10^2). The original estimator and its two extensions use the same candidate solutions. **Bold** and underlined values indicate the best and second-best results, respectively, for each base estimator, model, and metric.

Method	Qwen3-8B			Gemma-3-12B			Llama-3.1-8B		
	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑
CoT	22.57	21.86	70.01	24.10	22.93	62.98	25.57	28.15	63.06
CoT + RC	17.95	18.09	73.59	17.00	17.76	67.00	**19.08**	**21.11**	67.84
CoT + RC-A	**14.03**	**16.16**	**73.62**	**12.52**	**15.53**	**67.00**	20.45	21.30	**67.89**
SteerConf	15.29	16.49	75.76	19.05	19.17	67.34	23.83	25.42	69.54
SteerConf + RC	11.99	14.52	**76.26**	13.05	15.99	67.40	**17.80**	**20.10**	**70.41**
SteerConf + RC-A	**10.11**	**13.61**	76.24	**9.80**	**14.83**	**67.49**	19.30	20.53	70.26

6 Conclusion

This work introduces rubric-based calibration for verbalized confidence, complementing the model's holistic judgment with assessments under progressively stricter verification requirements. RC combines these assessments with direct confidence, while RC-A applies an additional correction based on profile repeatability and RC confidence. Experiments across eight benchmarks and three model families, together with analyses across model sizes, show that both methods improve calibration while maintaining competitive discrimination between correct and incorrect responses. Further analyses indicate that response-specific rubric scores provide information beyond simply producing more conservative confidence estimates and that their benefits extend to other confidence estimators. These findings highlight the value of explicitly specifying the requirements a response should satisfy for its reported confidence to be warranted, complementing existing efforts to improve how confidence is elicited and aggregated.

AI USE STATEMENT

The authors used AI tools to assist with language editing and improve the clarity and readability of the paper. All AI-assisted text was reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of this work.

REPRODUCIBILITY STATEMENT

The paper provides the information needed to reproduce the proposed methods and experiments. Sections 3 and 4 specify the methods and evaluation protocol, while the appendices provide the prompts, implementation details, and complete results. All experiments use publicly available models and benchmarks, facilitating independent verification.

REFERENCES

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 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)*, pp. 6577–6595, 2024.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783v1.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026.

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

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Yihan Hong, Huaiyuan Yao, Bolin Shen, Wanpeng Xu, Hua Wei, and Yushun Dong. RULERS: Locked rubrics and evidence-anchored scoring for robust LLM evaluation. *arXiv e-prints*, 2026. URL https://arxiv.org/abs/2601.08654v1.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022. URL https://arxiv.org/abs/2205.14334.

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

MathArena. AIME 2025. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/MathArena/aime_2025.

Ning Miao, Yee Whye Teh, and Tom Rainforth. SelfCheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *International Conference on Learning Representations*, volume 2024, pp. 56556–56571, 2024.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 2901–2907, 2015.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. URL https://arxiv.org/abs/2311.12022.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

KiJung Seo, Sehun Lim, and Taeuk Kim. ADVICE: Answer-dependent verbalized confidence estimation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23942–23962, 2026.

William F Shen, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. Rethinking rubric generation for improving LLM judge and reward modeling for open-ended tasks. *arXiv preprint arXiv:2602.05125*, 2026. URL https://arxiv.org/abs/2602.05125.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38, 2025.

Clemencia Siro, Pourya Aliannejadi, and Mohammad Aliannejadi. Learning to Judge: LLMs designing and applying evaluation rubrics. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6371–6389, 2026.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed H Chi, Denny Zhou, et al. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 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*, pp. 5433–5442, 2023.

Victor Wang and Elias Stengel-Eskin. Calibrating verbalized confidence with self-generated distractors. In *International Conference on Learning Representations*, volume 2026, pp. 121268–121297, 2026.

Yuxi Xia, Dennis Ulmer, Terra Blevins, Yihong Liu, Hinrich Schütze, and Benjamin Roth. Calibration is not enough: Evaluating confidence estimation under language variations. *arXiv preprint arXiv:2601.08064*, 2026. URL https://arxiv.org/abs/2601.08064.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, et al. RubricBench: Aligning model-generated rubrics with human standards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31179–31200, 2026.

Ziang Zhou, Tianyuan Jin, Jieming Shi, and Qing Li. SteerConf: Steering LLMs for confidence elicitation. *Advances in Neural Information Processing Systems*, 38:19460–19486, 2025.

A Experimental Details

A.1 Benchmark Details

To evaluate the reliability of verbalized confidence across different tasks, the experiments use eight benchmarks spanning mathematical reasoning, symbolic reasoning, commonsense reasoning, scientific reasoning, professional knowledge, and ethical judgment, with 3,910 questions in total. The tasks and corresponding datasets are described below.

- **Mathematical Reasoning:** GSM8K (Cobbe et al., 2021) contains grade-school mathematics word problems requiring multiple steps of calculation, assessing the ability to translate natural-language problems into mathematical reasoning. AIME25 (MathArena, 2025) contains more challenging competition mathematics problems that demand reasoning under complex constraints. The two benchmarks comprise 1,319 and 30 questions, respectively.
- **Symbolic Reasoning:** Date Understanding and Object Counting from BIG-Bench Hard (Suzgun et al., 2023) assess the ability to infer dates from temporal relations and to identify and count specified objects in text, respectively. Each benchmark contains 250 questions.
- **Commonsense Reasoning:** StrategyQA (Geva et al., 2021) contains 229 yes/no questions that typically require background knowledge and multiple reasoning steps, rather than answers directly available from the question wording.
- **Scientific Reasoning:** GPQA Diamond (Rein et al., 2023) contains 198 graduate-level science questions covering physics, chemistry, and biology. It examines confidence estimation on problems requiring specialized scientific knowledge and complex reasoning.
- **Professional Knowledge:** Professional Law from MMLU (Hendrycks et al., 2021) contains 1,534 questions requiring an understanding of legal knowledge and its application to specific situations.
- **Ethical Judgment:** Business Ethics from MMLU (Hendrycks et al., 2021) contains 100 questions assessing the understanding and application of ethical principles in business decision-making.

A.2 Generation and Evaluation Settings

Response generation. Each model first generates one candidate solution per question. The prompt is adapted to the task and requires step-by-step reasoning. All models use greedy decoding with a maximum of 2,048 output tokens. Qwen3's separate thinking mode is disabled for both response generation and confidence elicitation. Once generated, the candidate is held fixed across confidence estimators for that model, without generating a new solution for each estimator. Table 4 reports accuracy on each benchmark.

Table 4: Candidate solution accuracy (%) on the eight benchmarks. Each model is evaluated on all N questions.

Dataset	N	Qwen3-8B	Qwen3-14B	Qwen3-32B	Gemma-3-12B	Llama-3.1-8B
GSM8K	1,319	93.40	95.45	95.07	93.93	84.08
AIME25	30	23.33	20.00	20.00	16.67	0.00
Date Understanding	250	88.40	92.00	90.00	94.00	66.00
Object Counting	250	96.00	92.40	97.60	96.00	71.60
StrategyQA	229	66.81	71.18	73.36	72.49	65.50
GPQA Diamond	198	36.87	45.96	53.03	35.86	19.70
Professional Law	1,534	53.19	58.02	62.91	54.89	49.87
Business Ethics	100	78.00	81.00	82.00	71.00	71.00

Confidence elicitation. The evaluation input contains the original question and the full candidate solution, and prompts request numerical confidence scores from 0 to 100. All methods except Sampling use greedy decoding, with a maximum of 2,048 output tokens per call. Sampling performs five stochastic confidence evaluations of the same candidate, with temperature 0.7 and top-p 1.0.

Method configuration. RC and RC-A use the same five-level rubric and four English prompt paraphrases across all models, yielding four direct confidence estimates and four sets of rubric scores per response. The two methods reuse the same model outputs and differ only in aggregation. RC-A therefore requires no additional model calls, and each method requires eight evaluator calls per response. Appendix C.4 compares these call budgets with those of the baselines.

A.3 Baseline Implementations

Direct. Given the problem and candidate solution, the model reports the probability that the candidate is fully correct without providing an analysis (Tian et al., 2023). This reported value is used as the confidence estimate.

CoT. In a single call, the model examines the candidate's reasoning, calculations, factual claims, and final answer step by step before reporting its confidence (Kojima et al., 2022; Xiong et al., 2024).

Multi-Step. Following the multi-step prompting strategy of Xiong et al. (2024), the model divides the candidate solution into logical steps, assesses each step, and reports both step-level confidence and final confidence. The final confidence serves as the estimate.

Self-Probing. Following the self-probing strategy of Xiong et al. (2024), this baseline evaluates the fixed candidate in two calls. The first asks the model to identify errors or weaknesses without reporting confidence. The second presents this analysis together with the original problem and candidate, then requests a confidence estimate.

Sampling. Following self-random sampling with confidence averaging (Xiong et al., 2024), the Direct prompt elicits five stochastic confidence judgments of the same candidate solution. Their arithmetic mean is used as the final estimate.

Paraphrase Mean. To control for prompt wording, the original Direct prompt and four semantically equivalent English paraphrases evaluate the same candidate while preserving the evaluation objective. The resulting five confidence estimates are averaged.

Multilingual Mean. To control for elicitation language, the Direct prompt, including the system instruction, is translated into five languages while the problem and candidate solution remain in their original form. The resulting confidence estimates are averaged. Qwen and Gemma use English, Chinese, Japanese, French, and Spanish. Llama uses English, German, French, Italian, and Spanish.

SteerConf. Following Zhou et al. (2025), the model reports verbalized confidence under five steering conditions: very cautious, cautious, neutral, confident, and very confident. These scores are aggregated using their mean and consistency across conditions. Because every condition evaluates the same fixed candidate solution, the answer-consistency factor is set to one.

A.4 Metrics

Let $c_i \in [0, 1]$ denote the confidence assigned to response i and $y_i \in \{0, 1\}$ its correctness label. For each benchmark, metrics are computed over the N responses for which the evaluated method produces valid confidence estimates. We additionally repeat the comparison on the intersection of responses with valid confidence estimates from all methods. The results of this common-sample comparison are summarized in Appendix B.1.

Expected calibration error (ECE). ECE measures the discrepancy between confidence and accuracy within confidence bins (Naeini et al., 2015). The interval $[0, 1]$ is divided into ten equal-width bins, $[0, 0.1), \ldots, [0.9, 1]$. Let B_m contain the responses in bin m. Then

$$\mathrm{ECE} = \sum_{m:\,|B_m|>0} \frac{|B_m|}{N} \left| \frac{1}{|B_m|} \sum_{i \in B_m} c_i - \frac{1}{|B_m|} \sum_{i \in B_m} y_i \right|. \tag{5}$$

A lower ECE therefore indicates closer agreement between reported confidence and empirical accuracy.

Brier score. The Brier score (Brier, 1950) measures the mean squared error of confidence against the correctness labels:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(c_i - y_i)^2. \tag{6}$$

A lower Brier score indicates a smaller squared deviation from the correctness labels.

AUROC. The area under the receiver operating characteristic curve measures how well confidence separates correct from incorrect responses (Fawcett, 2006). Equivalently, it is the probability that a randomly chosen correct response receives higher confidence than a randomly chosen incorrect response, with ties receiving half credit. A higher AUROC therefore indicates better discrimination.

PR-P and PR-N. PR-P and PR-N assess precision–recall performance (Saito & Rehmsmeier, 2015). Both are computed using average precision, $\text{AP} = \sum_j (r_j - r_{j-1})p_j$, where p_j and r_j are precision and recall after including all responses at the jth distinct score threshold, in descending order, with $r_0 = 0$. PR-P treats correct responses as positives and ranks by c_i. PR-N treats incorrect responses as positives and ranks by $1 - c_i$. Higher PR-P and PR-N indicate better detection of their respective positive classes.

A.5 Prompts for Rubric Calibration

The original direct confidence and rubric scoring prompts are shown below. For each elicitation type, three additional English prompt paraphrases are generated by an LLM, giving four prompt paraphrases while preserving the evaluation requirements. RC and RC-A use the same prompts and model outputs.

System instruction. You are a careful evaluator of candidate solutions.

Direct confidence (origin)

Problem:
{question}

Candidate solution:
{candidate_solution}

Estimate the probability that the candidate solution is fully correct. Evaluate the provided candidate solution rather than producing a new solution.

Return only the following JSON object:
{"confidence": <number from 0 to 100>}

The confidence must be a numerical value from 0 to 100. Do not include a percent sign. Do not include any analysis, explanation, or additional text.

Rubric scoring (origin)

Problem:
{question}

Candidate solution:
{candidate_solution}

Evaluate the candidate solution exactly as provided rather than producing, repairing, or substituting another solution.

Score the same candidate solution under five increasingly strict verification levels. The levels are not separate aspects; they are a difficulty ladder for deciding whether the solution can still be accepted as correct under stronger scrutiny.
1. level_1_surface_validity: Surface validity: The solution is relevant, has a plausible final answer form, and contains no immediately obvious nonsense.
2. level_2_local_correctness: Local correctness: The local calculations, factual claims, option references, and small inference steps are individually correct.
3. level_3_logical_sufficiency: Logical sufficiency: The stated reasoning is sufficient to support the conclusion, without decisive gaps or unjustified leaps.
4. level_4_adversarial_robustness: Adversarial robustness: The solution survives checks for hidden constraints, edge cases, counterexamples, misleading wording, units, and exception cases.
5. level_5_full_verification: Full verification: Under the strictest standard, the entire solution is fully correct and the final answer is verifiably entailed by the problem.

Higher levels are stricter. A high score at a stricter level should mean the solution would also pass easier levels.

Return only the following JSON object:
{
"level_1_surface_validity": "<number from 0 to 100>",
"level_2_local_correctness": "<number from 0 to 100>",
"level_3_logical_sufficiency": "<number from 0 to 100>",
"level_4_adversarial_robustness": "<number from 0 to 100>",
"level_5_full_verification": "<number from 0 to 100>"
}

Each value must be a numerical probability from 0 to 100 that the candidate solution passes that verification level. Do not include a percent sign. Do not output an overall or final confidence. Do not include any analysis, explanation, or additional text.

B Additional Experimental Results

B.1 Detailed Main Results

Figures 6, 7, and 8 present per-benchmark comparisons of RC, RC-A, and the eight baselines across all five metrics for Qwen3-8B, Gemma-3-12B, and Llama-3.1-8B, respectively. All results use the same candidate correctness labels as Table 1. When a metric is undefined, the corresponding entry is marked N/A rather than plotted as zero and is omitted only from that metric's macro average. On AIME25, Llama-3.1-8B has no correct responses. For Gemma-3-12B on AIME25, Multi-Step produces valid confidence estimates for 12 of the 30 responses, none of which are correct. Consequently, its AUROC and PR-P macro averages include seven benchmarks, whereas ECE, Brier, and PR-N include all eight.

Common-sample comparison. To assess whether differences in sample coverage affect the main findings, we repeat the comparison using only responses with valid confidence estimates from all methods. RC-A retains the lowest macro-average ECE and Brier across all three models, supporting the calibration gains observed in the main comparison.

Qwen3-8B. RC-A improves calibration relative to RC across most tasks, reducing ECE on all eight benchmarks and Brier on seven. It also achieves the lowest ECE and Brier among the compared methods on AIME25, StrategyQA, GPQA Diamond, and Business Ethics. The discrimination results are more task-dependent. Both variants perform well on GSM8K and Business Ethics, whereas their AUROC remains below several baselines on the two symbolic reasoning benchmarks.

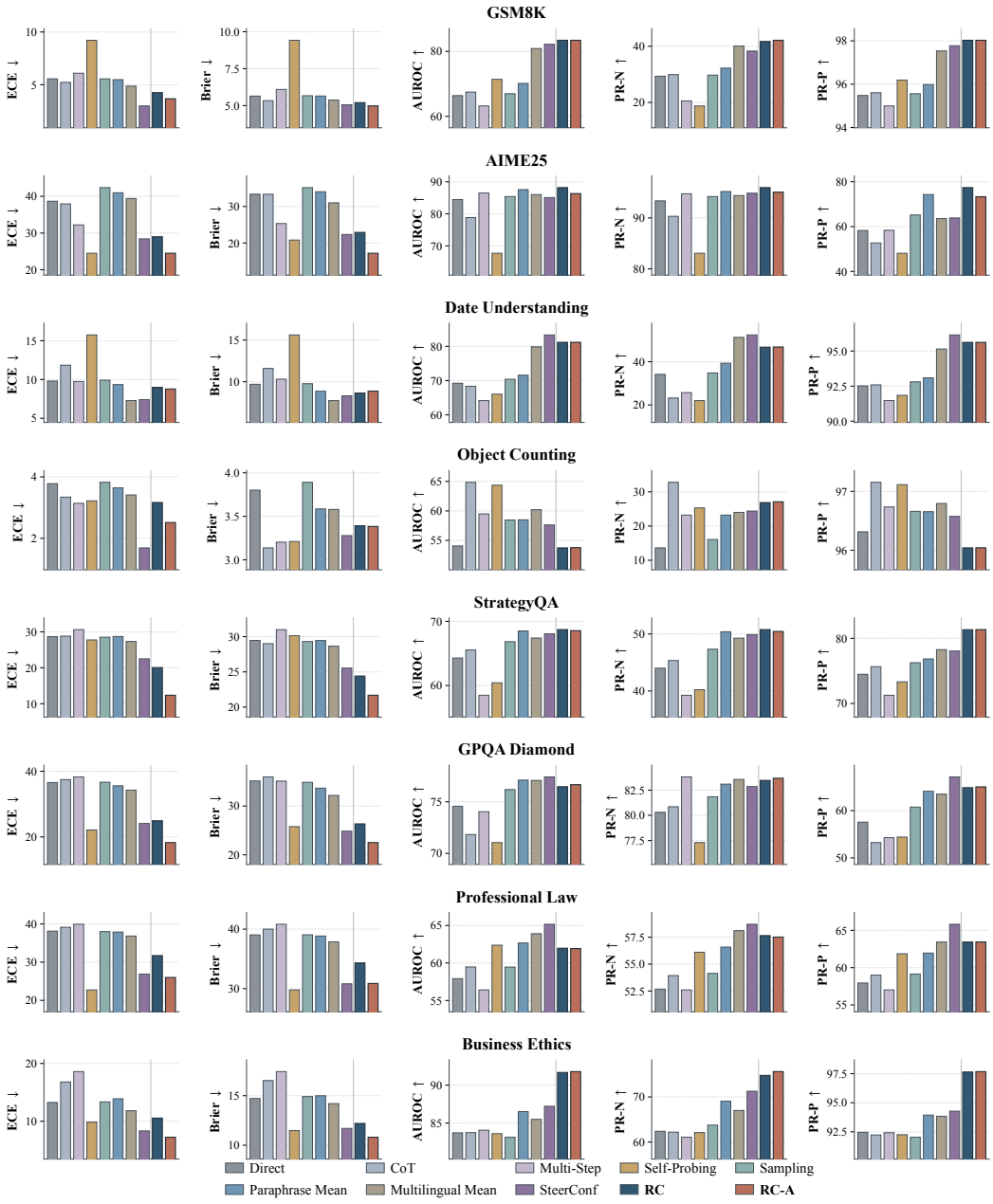


Figure 6: Per-benchmark results for Qwen3-8B. Rows correspond to benchmarks; columns show ECE, Brier, AUROC, PR-N, and PR-P. All metrics are scaled by 10^2. Bars follow the legend order, with RC and RC-A shown last.

Gemma-3-12B. RC-A attains the lowest ECE and Brier on StrategyQA, GPQA Diamond, Professional Law, and Business Ethics. Relative to RC, its additional correction substantially lowers ECE on AIME25 and Professional Law, while the two methods yield similar results on Object Counting. On Date Understanding, however, RC-A increases both ECE and Brier. As with Qwen3-8B, discrimination remains task-dependent.

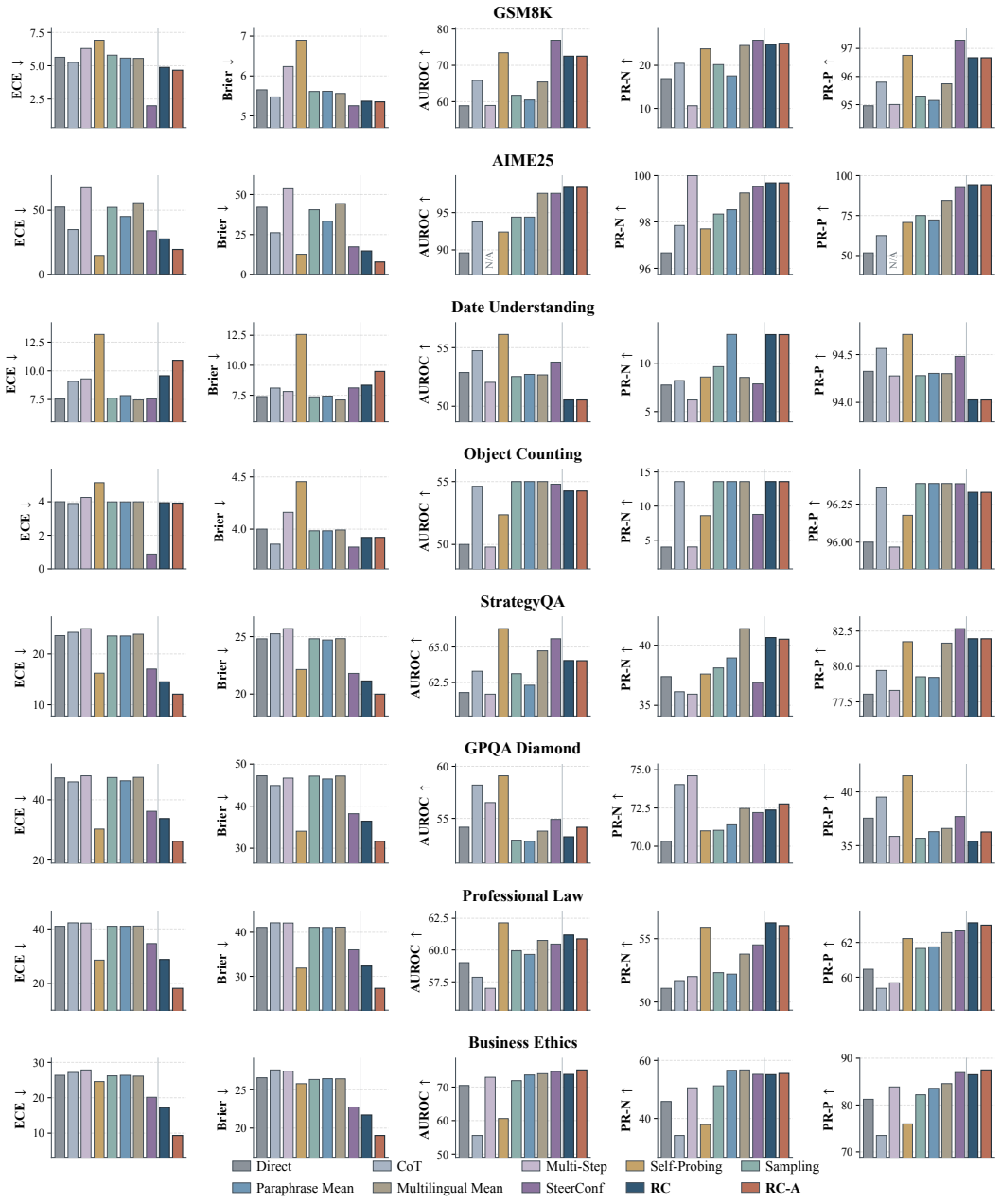


Figure 7: Per-benchmark results for Gemma-3-12B. Rows correspond to benchmarks; columns show ECE, Brier, AUROC, PR-N, and PR-P. All metrics are scaled by 10^2. Bars follow the legend order, with RC and RC-A shown last. N/A marks an undefined metric, not a zero score.

Llama-3.1-8B. The task-level trade-offs are more pronounced for this model. RC-A achieves the lowest Brier on AIME25, GPQA Diamond, and Professional Law, together with the highest AUROC on Date Understanding among the compared methods. However, it increases ECE and Brier relative to RC on GSM8K, StrategyQA, and Business Ethics. The macro-average gains therefore coexist with tasks on which the simpler RC estimator provides better calibration.

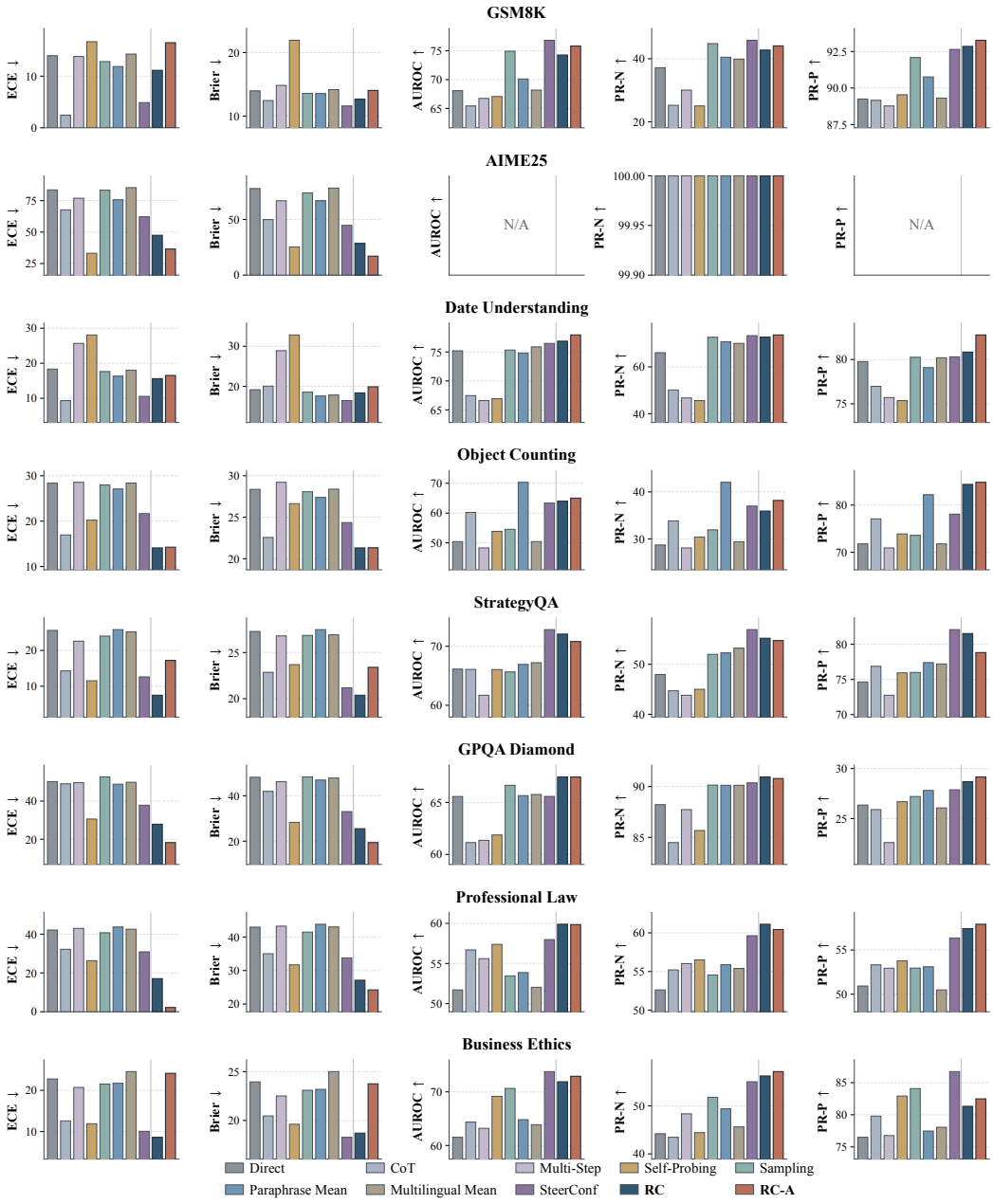


Figure 8: Per-benchmark results for Llama-3.1-8B. Rows correspond to benchmarks; columns show ECE, Brier, AUROC, PR-N, and PR-P. All metrics are scaled by 10^2. Bars follow the legend order, with RC and RC-A shown last. N/A marks an undefined metric, not a zero score.

B.2 Complete Ablation Results

Table 5 extends the ablations in Section 4.3 to all five metrics, comparing Direct Mean, Rubric Mean, RC, the two single-condition variants, and RC-A. All variants reuse the same candidate solutions and elicited scores, with both RC-A thresholds fixed at 0.9. In addition to ECE, Brier, and AUROC, the table reports PR-N and PR-P to examine how the components affect the identification of incorrect and correct responses.

Table 5: Complete component ablations across eight benchmarks (macro averages; all metrics are scaled by 10^2). The variants and thresholds match Table 2. **Bold** and underlined values indicate the best and second-best results, respectively, for each model and metric.

Metric	Direct Mean	Rubric Mean	RC	A_R only	C_{RC} only	RC-A
			Qwen3-8B			
ECE ↓	22.14	22.56	16.57	13.10	15.93	**12.91**
Brier ↓	21.37	21.20	17.17	15.11	16.96	**15.04**
AUROC ↑	71.71	75.21	**75.67**	75.32	74.94	75.45
PR-N ↑	54.60	**60.37**	59.73	59.68	59.19	59.79
PR-P ↑	81.33	83.02	**84.29**	83.52	83.48	83.81
			Gemma-3-12B			
ECE ↓	25.06	21.75	17.54	13.40	17.04	**13.12**
Brier ↓	23.72	20.65	18.00	15.67	17.83	**15.58**
AUROC ↑	63.85	65.80	65.99	66.22	65.13	**66.24**
PR-N ↑	44.90	45.01	46.93	**47.03**	45.85	47.02
PR-P ↑	77.31	80.90	81.03	81.21	80.72	**81.25**
			Llama-3.1-8B			
ECE ↓	33.42	21.86	18.66	18.53	18.33	**18.19**
Brier ↓	32.98	23.40	21.57	20.40	21.54	**20.37**
AUROC ↑	66.71	68.89	69.49	69.79	69.66	**69.97**
PR-N ↑	62.45	63.53	64.35	64.82	64.40	**64.87**
PR-P ↑	69.66	72.17	72.43	72.44	72.73	**72.75**

Combining direct confidence with rubric scores, RC reduces both ECE and Brier on all three models and improves AUROC over Direct Mean and Rubric Mean. This shows that the two sources of information improve calibration while strengthening the distinction between correct and incorrect responses. Building on RC, RC-A further reduces ECE and Brier on all three models, improves AUROC on Gemma-3-12B and Llama-3.1-8B, and improves PR-N on all three models. On Qwen3-8B, RC retains slightly higher AUROC and PR-P than RC-A.

Compared with the two single-condition variants, RC-A performs better in ECE, Brier, AUROC, and PR-P on all three models. For PR-N, it performs better on Qwen3-8B and Llama-3.1-8B and nearly matches the A_R-only variant on Gemma-3-12B (47.02 versus 47.03). Overall, RC provides consistent gains by incorporating rubric information, while RC-A further improves calibration and maintains strong discrimination.

Comparison with unconditional correction. To isolate the role of the adaptive gate, we compare RC-A with $C_{\mathrm{uncond}} = D_{\mathrm{mean}} R^2_{\mathrm{mean}}$, which applies the additional rubric factor to every response. The comparison covers 11,730 candidate responses generated by Qwen3-8B, Gemma-3-12B, and Llama-3.1-8B, with equal weight assigned to each model's benchmark-macro results. Table 6 shows that, at the default thresholds and the reported neighboring configurations, RC-A outperforms unconditional correction on all five metrics, indicating that the gains are not confined to a single parameter setting. These results support using the gate to apply additional rubric correction selectively, improving overall calibration and discrimination while avoiding unnecessary confidence reductions.

Table 6: Adaptive versus unconditional rubric correction (equal-weight averages of the three models' benchmark-macro results; all metrics are scaled by 10^2). The default thresholds are $\lambda_a = \lambda_b = 0.9$. **Bold** and underlined values indicate the best and second-best results, respectively, for each metric.

Method	λ_a	λ_b	ECE ↓	Brier ↓	AUROC ↑	PR-N ↑	PR-P ↑
Unconditional	–	–	15.41	17.28	70.28	57.00	79.10
RC-A	0.80	0.80	15.12	17.13	70.40	57.14	79.12
RC-A	0.85	0.85	14.76	**16.95**	70.42	57.25	79.25
RC-A	0.85	0.90	14.84	17.00	70.53	**57.30**	79.24
RC-A	0.90	0.85	**14.65**	16.96	70.39	57.15	79.22
RC-A (default)	0.90	0.90	14.74	17.00	**70.55**	57.22	**79.27**

B.3 Results Across Model Sizes

Table 7 supplements the model size analysis in Section 5.2 with complete results for Qwen3-8B, Qwen3-14B, and Qwen3-32B across eight benchmarks. In addition to ECE, Brier, and AUROC, it reports PR-N and PR-P for all baselines, RC, and RC-A. The verification requirements and aggregation rules remain unchanged across the three sizes, with both RC-A thresholds fixed at 0.9 rather than adjusted for each scale.

Table 7: Complete model size comparisons across eight benchmarks (macro averages; all metrics are scaled by 10^2). **Bold** and underlined values indicate the best and second-best results, respectively, for each model and metric. The 8B results reproduce Table 1.

Metric	Direct	CoT	Multi-Step	Self-Probing	Sampling	Paraphrase Mean	Multilingual Mean	SteerConf	RC	RC-A
Qwen3-8B										
ECE ↓	21.80	22.57	22.32	16.89	22.27	21.93	20.65	15.29	16.57	**12.91**
Brier ↓	21.36	21.86	21.16	18.28	21.58	21.14	20.07	16.49	17.17	**15.04**
AUROC ↑	69.33	70.01	68.29	68.37	70.85	72.82	75.11	**75.76**	75.67	75.45
PR-N ↑	51.23	52.32	50.12	48.11	52.73	56.13	58.45	59.07	59.73	**59.79**
PR-P ↑	78.12	77.27	77.06	76.88	79.80	82.11	81.51	82.45	**84.29**	83.81
Qwen3-14B										
ECE ↓	22.73	21.55	21.47	14.85	22.51	22.67	23.24	19.23	13.14	**9.35**
Brier ↓	21.36	19.00	19.40	15.02	21.11	20.98	21.83	18.26	14.16	**12.65**
AUROC ↑	65.98	69.19	67.24	72.08	70.88	72.29	72.88	74.61	**76.44**	76.22
PR-N ↑	43.40	47.49	45.55	48.03	48.16	48.20	49.13	51.12	**53.20**	53.01
PR-P ↑	79.23	85.23	82.41	87.28	83.66	86.75	86.29	87.44	**89.25**	89.07
Qwen3-32B										
ECE ↓	20.05	20.94	20.33	15.72	20.70	20.19	20.41	17.54	12.67	**9.06**
Brier ↓	19.34	19.53	19.73	15.81	19.69	19.07	19.56	17.49	13.88	**12.52**
AUROC ↑	66.31	63.87	65.64	67.31	72.46	69.66	69.91	70.06	**72.70**	72.65
PR-N ↑	40.50	38.85	40.93	42.08	45.74	44.17	44.07	45.37	48.83	**49.40**
PR-P ↑	83.52	81.06	81.19	82.21	87.07	87.49	87.64	85.07	**88.71**	88.54

RC-A achieves the lowest ECE and Brier at all three sizes, while RC and RC-A obtain the two best results on all five metrics at 14B and 32B. The additional discrimination metrics reveal a similar trade-off to AUROC: RC retains higher PR-P than RC-A at every size, whereas RC-A improves PR-N at 8B and 32B but slightly reduces it at 14B. Overall, RC and RC-A exhibit strong scalability under a shared configuration, improving calibration while maintaining strong discrimination.

B.4 Extensions to Other Confidence Estimators

Table 8 supplements Section 5.3 with complete results for the rubric extensions of CoT and SteerConf. Both extensions reuse the verification requirements and rubric scores from the main experiments, without redesigning the prompts or retuning the RC-A thresholds for the base estimator.

Table 8: Complete rubric-extension results across eight benchmarks (macro averages; all metrics are scaled by 10^2). **Bold** and underlined values indicate the best and second-best results, respectively, for each base estimator, model, and metric.

	CoT			SteerConf		
Metric	Base	+ RC	+ RC-A	Base	+ RC	+ RC-A
Qwen3-8B						
ECE ↓	22.57	17.95	**14.03**	15.29	11.99	**10.11**
Brier ↓	21.86	18.09	**16.16**	16.49	14.52	**13.61**
AUROC ↑	70.01	73.59	**73.62**	75.76	**76.26**	76.24
PR-N ↑	52.32	59.15	**59.17**	59.07	60.77	**60.83**
PR-P ↑	77.27	**79.83**	79.78	82.45	**83.82**	83.63
Gemma-3-12B						
ECE ↓	24.10	17.00	**12.52**	19.05	13.05	**9.80**
Brier ↓	22.93	17.76	**15.53**	19.17	15.99	**14.83**
AUROC ↑	62.98	67.00	**67.00**	67.34	67.40	**67.49**
PR-N ↑	42.02	45.75	**46.01**	45.10	46.43	**46.51**
PR-P ↑	75.16	**81.56**	81.47	81.33	**81.60**	81.52
Llama-3.1-8B						
ECE ↓	25.57	**19.08**	20.45	23.83	**17.80**	19.30
Brier ↓	28.15	**21.11**	21.30	25.42	**20.10**	20.53
AUROC ↑	63.06	67.84	**67.89**	69.54	**70.41**	70.26
PR-N ↑	54.63	61.82	**62.51**	64.76	65.65	**65.75**
PR-P ↑	68.44	**72.22**	72.02	72.01	**73.29**	72.99

SteerConf retains its original confidence aggregation, while the rubric prompts do not incorporate its steering scenarios.

For a base confidence B, the RC extension uses $C_{\mathrm{RC},B} = BR_{\mathrm{mean}}$. The RC-A extension additionally multiplies $C_{\mathrm{RC},B}$ by R_{mean} when $A_R \geq 0.9$ or $C_{\mathrm{RC},B} \geq 0.9$, applying this extra factor only once when both conditions hold. Each base estimator and its two extensions are evaluated on the same candidate solutions and samples with valid confidence estimates. This ensures identical sample coverage across the three variants.

Both rubric extensions improve all five metrics over their respective base estimators on all three models. Beyond this initial gain, the adaptive gate further reduces ECE and Brier on Qwen3-8B and Gemma-3-12B, but increases both relative to the RC extension on Llama-3.1-8B. Across all six model–estimator combinations, the adaptive gate improves PR-N while slightly reducing PR-P. These results show that rubric information is not limited to direct confidence estimation. Without modifying the prompting or aggregation of the base estimator, rubric calibration improves calibration and discrimination across models.

C Parameter Sensitivity and Inference Budget

C.1 Threshold Sensitivity

RC-A uses two thresholds to identify responses for additional correction. The analysis varies either λ_a or λ_b from 0 to 1 in increments of 0.01, while fixing the other at 0.9. Figure 9 displays the range $[0.85, 1]$ for Qwen3-8B and Gemma-3-12B, and $[0.5, 1]$ for Llama-3.1-8B. All comparisons reuse the same candidate solutions and elicited scores, with the OR rule and single additional rubric factor unchanged.

The finer grid reveals model-specific differences in local sensitivity. The Qwen3-8B results remain locally stable around the default configuration, with each metric varying by less than 0.1 on the displayed scale when either threshold ranges from 0.85 to 0.95 and the other is fixed at 0.9. Its calibration errors increase more sharply as λ_a approaches 1. Gemma-3-12B exhibits a similar pattern: ECE and Brier generally increase when λ_a rises above 0.9, whereas changes in λ_b have a smaller

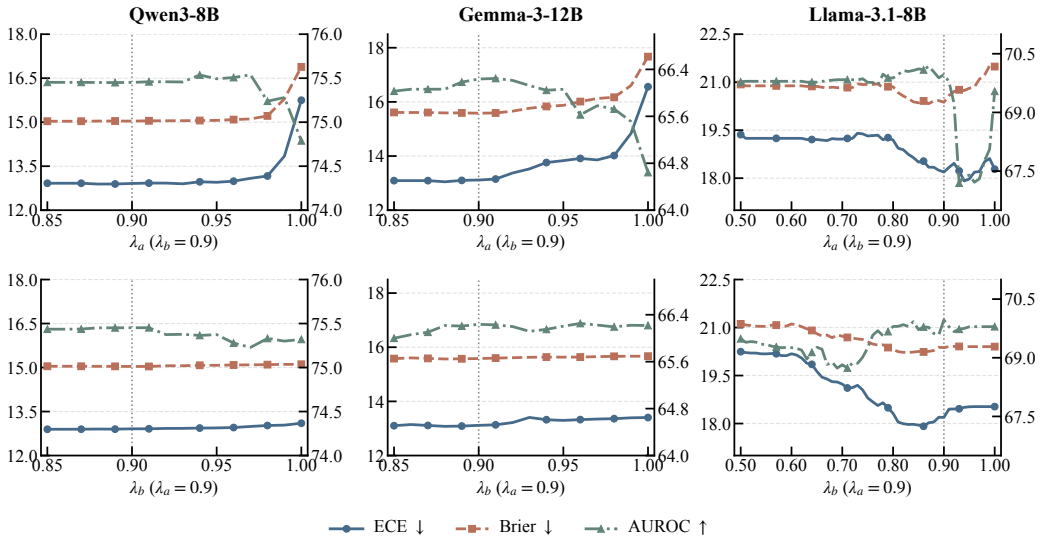


Figure 9: Sensitivity to the two RC-A thresholds across eight benchmarks (macro averages; all metrics are scaled by 10^2). The top row varies λ_a with $\lambda_b = 0.9$. The bottom row varies λ_b with $\lambda_a = 0.9$. ECE and Brier use the left axis, while AUROC uses the right axis. Dotted vertical lines mark the default threshold of 0.9.

effect. Llama-3.1-8B shows the greatest local sensitivity, particularly to λ_a. For example, increasing λ_a from 0.92 to 0.93 reduces AUROC from 69.23 to 67.19. At $\lambda_a = 0.94$, ECE is lower than at the default setting (17.92 versus 18.19), but Brier is higher (20.74 versus 20.37) and AUROC is lower (67.36 versus 69.97). These results indicate that Qwen3-8B and Gemma-3-12B have a wider stable threshold region around the default setting, whereas Llama-3.1-8B is more sensitive to threshold changes.

C.2 Number of Prompt Paraphrases

The number of prompt paraphrases controls both the available assessment information and the evaluation budget. This analysis compares RC and RC-A using $V \in \{1, 2, 3, 4\}$ paraphrases, with one direct confidence elicitation and one five-level rubric elicitation per paraphrase, requiring $2V$ calls per candidate solution. For each V, all $\binom{4}{V}$ subsets of the four stored prompt indices are evaluated, giving four, six, four, and one subsets, respectively. Each index subset is applied separately within the direct and rubric elicitation sets; the shared index labels are used only for bookkeeping and do not imply semantic pairing between their prompts. Each subset then recomputes the corresponding means and profile repeatability. Both thresholds remain fixed at 0.9.

For $V = 1$, profile repeatability across paraphrases is undefined. The RC-A variant therefore disables the A_R condition and retains only the $C_{\mathrm{RC}} \geq 0.9$ condition; this convention does not interpret a single profile as having low repeatability. Consequently, the change from one to multiple paraphrases also enables the repeatability condition, rather than reflecting averaging alone. For each subset, metrics are first calculated within each benchmark and then macro-averaged. Figure 10 shows the mean and range of these macro scores across equal-size subsets.

The subset-averaged Brier score decreases as more paraphrases are included for both methods on all three models. The gains are particularly clear on Llama-3.1-8B: from two to four paraphrases, RC-A's Brier decreases from 21.35 to 20.37 and AUROC increases from 67.16 to 69.97. Gemma-3-12B also benefits in AUROC, which rises from 65.74 to 66.24 over the same range. The subset-averaged PR-N and PR-P also increase with paraphrase count for both methods on all three models. However, neither ECE nor AUROC improves monotonically in every setting. On Qwen3-8B, RC-A's mean AUROC is slightly higher with two paraphrases than with four (75.72 versus 75.45), while Gemma-3-12B's mean ECE is lower with two paraphrases (12.94 versus 13.12). Thus, multiple paraphrases can improve calibration and discrimination, but four paraphrases are not uniformly optimal across metrics; the appropriate budget depends on the model and the desired trade-off.

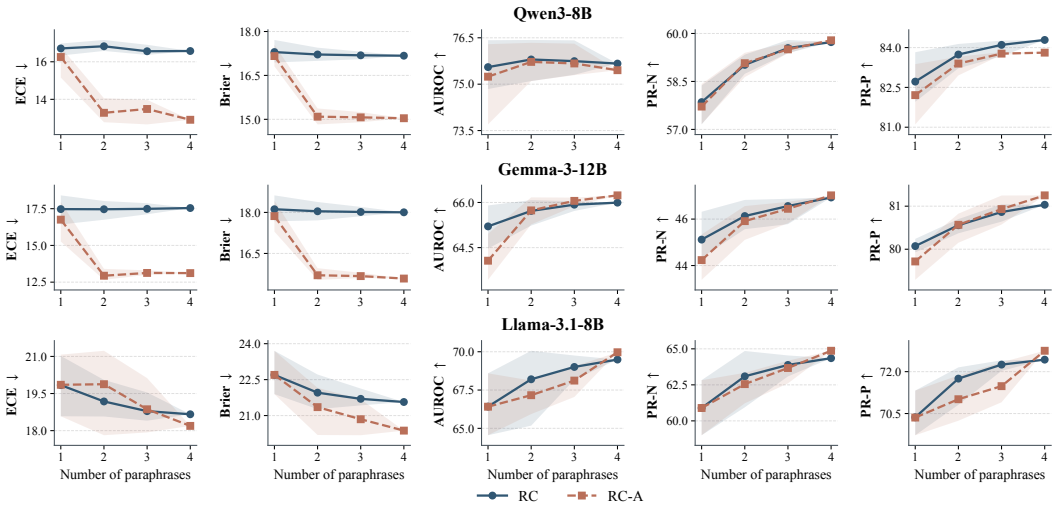


Figure 10: Effect of the number of prompt paraphrases on RC and RC-A across eight benchmarks (macro averages; all metrics are scaled by 10^2). Rows correspond to models; columns show ECE, Brier, AUROC, PR-N, and PR-P. Lines show means across all same-size subsets, and shaded regions show their minimum and maximum. The single-paraphrase RC-A variant uses only the high-confidence condition.

C.3 Effect of Verification Levels

The preceding analysis varies the number of prompt paraphrases while keeping the rubric fixed. Here, the four paraphrases and the direct confidence elicitation remain unchanged, and different verification levels contribute to the rubric aggregate. To isolate this choice from the additional RC-A correction, the comparison uses RC and progressively includes the first K levels:

$$C_K = D_{\text{mean}}\left(\frac{1}{K}\sum_{k=1}^{K}\bar{R}_k\right), \qquad K \in \{1,2,3,4,5\}, \tag{7}$$

where $\bar{R}_k$ is the mean score for level k across the four paraphrases. All scores come from the existing five-level evaluations.

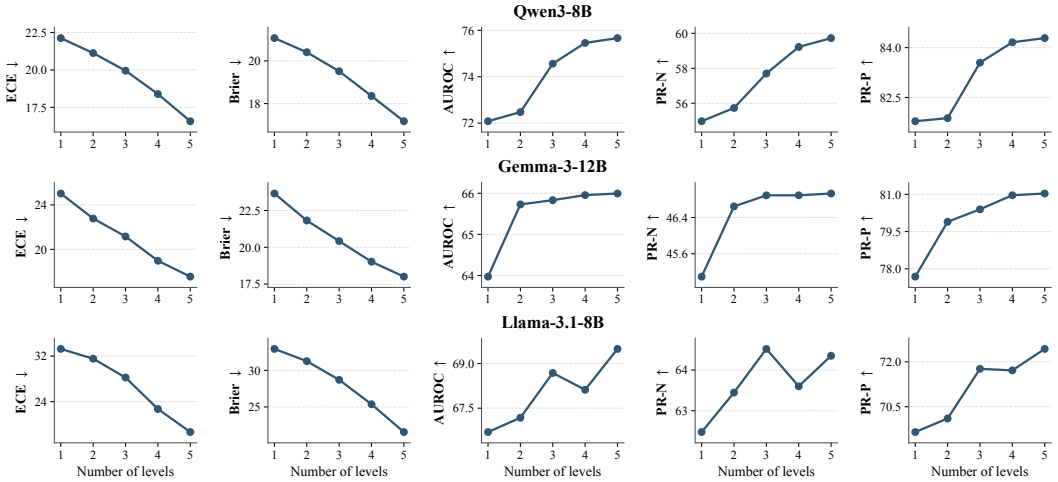


Figure 11: Effect of progressively including verification levels in RC across eight benchmarks (macro averages; all metrics are scaled by 10^2). Rows correspond to models; columns show ECE, Brier, AUROC, PR-N, and PR-P. At K levels, the rubric aggregate includes levels 1 through K. Four prompt paraphrases are used throughout.

Figure 11 shows that ECE and Brier decrease at every step for all three models as stricter verification levels are included. From one to five levels, ECE decreases from 22.13 to 16.57 on Qwen3-8B,

from 25.03 to 17.54 on Gemma-3-12B, and from 33.26 to 18.66 on Llama-3.1-8B. AUROC also improves progressively on Qwen and Gemma, while Llama exhibits a decrease at four levels before reaching its highest value at five. Overall, as the verification requirements become stricter, RC's calibration errors decrease consistently, and the complete five-level rubric provides stronger overall performance. Progressive verification therefore provides useful information for confidence calibration.

C.4 Inference Budget

The evaluation budget is counted per fixed candidate solution, excluding the response generation shared by all confidence estimators. Direct, CoT, and Multi-Step each use one evaluator call, while Self-Probing uses two. Sampling, Paraphrase Mean, Multilingual Mean, and SteerConf each use five. The default RC and RC-A configuration uses eight calls: four for direct confidence and four for rubric scoring. Verification scores at all five levels are obtained in a single rubric call. RC-A reuses exactly the same outputs as RC. Its additional correction therefore requires no further model calls.

The default configuration thus uses more evaluator calls than the baselines to combine direct judgments with rubric assessments. With V prompt paraphrases, both RC and RC-A require $2V$ calls; the effect of reducing this budget is examined in Appendix C.2.

D Additional Analysis

D.1 Motivating Observations on Gemma-3-12B

Figure 12 shows a similar pattern for Gemma-3-12B under the setting described in Section 3.1. The error bars in Figures 1 and 12 are 95% pointwise percentile bootstrap intervals from 5,000 resamples of whole responses within each correctness group. Panel (a) shows the direct-confidence distributions for correct and incorrect responses, whereas panel (c) shows the proportions of all responses in the same confidence categories after RC calibration, separated by correctness. Both panels display exact 1.00 separately. Panel (d) covers the full confidence range and omits empty bins. The results show that mean direct confidence is 0.980 for correct responses and 0.919 for incorrect ones. Among the 943 responses assigned 0.950 under all four prompt paraphrases, only 48.5% are correct. A further 149 incorrect responses receive 1.000 under every paraphrase. Thus, repeated high-confidence judgments can remain poorly calibrated.

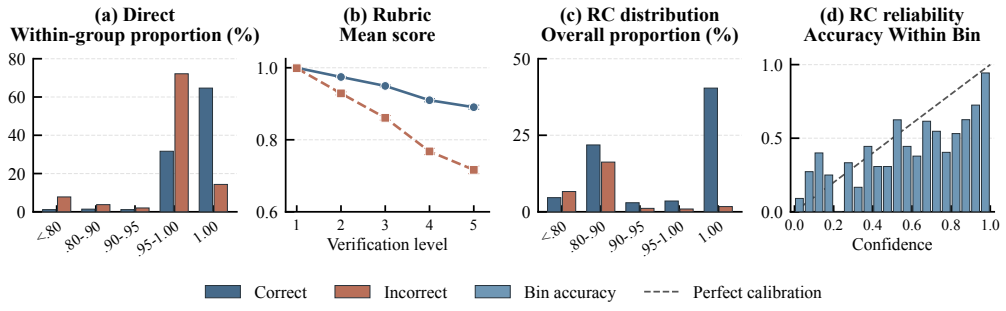


Figure 12: Gemma-3-12B on eight benchmarks. (a) Direct-confidence distributions within correct and incorrect responses. (b) Mean rubric scores, with 95% bootstrap confidence intervals; levels 1–5 correspond to the verification levels defined in Section 3.2. (c) Proportions of all responses across RC-calibrated confidence categories, separated by correctness. (d) Accuracy within RC confidence bins; the dashed diagonal indicates perfect calibration. Panel (d) uses bins of width 0.05.

By contrast, the rubric profiles show greater separation as verification requirements become stricter. At surface validity, both groups receive near-maximal scores, with a mean gap of 0.001. At full verification, the means are 0.890 for correct responses and 0.717 for incorrect ones, giving a gap of 0.174. The pooled gap widens at each successive level. This pattern also holds within seven of the eight benchmarks, with Object Counting as the exception. Panels (c–d) complement these

observations with the distribution of RC scores and the corresponding bin-wise accuracy on Gemma-3-12B.

D.2 Profile Repeatability and Residual Overconfidence

Using the adopted thresholds $\lambda_a = \lambda_b = 0.9$, responses are partitioned into four groups according to profile repeatability A_R and the RC confidence C_{RC}. Table 9 supplements Section 3.3 by comparing Direct Mean and RC within these same groups, pooling all 3,910 responses for each model. For a group g, let Acc_g denote its accuracy and $\langle\cdot\rangle_g$ the mean over its responses. The signed gaps are

$$\mathrm{Gap}_g(D) = \langle D_{\mathrm{mean}}\rangle_g - \mathrm{Acc}_g, \qquad \mathrm{Gap}_g(\mathrm{RC}) = \langle C_{\mathrm{RC}}\rangle_g - \mathrm{Acc}_g.$$

Positive and negative gaps indicate group-level overconfidence and underconfidence, respectively.

Table 9: Direct Mean and RC within groups defined by the adopted thresholds $\lambda_a = \lambda_b = 0.9$. Confidence and accuracy are reported on the $[0, 1]$ scale. Gap denotes mean confidence minus accuracy; dashes denote an empty group.

A_R	C_{RC}	Accuracy	Mean Direct	Mean RC	Gap (Direct)	Gap (RC)
			Qwen3-8B			
$< \lambda_a$	$< \lambda_b$	0.647	0.911	0.772	0.264	0.125
$< \lambda_a$	$\geq \lambda_b$	0.954	1.000	0.994	0.045	0.040
$\geq \lambda_a$	$< \lambda_b$	0.456	0.846	0.745	0.390	0.289
$\geq \lambda_a$	$\geq \lambda_b$	0.676	0.960	0.921	0.285	0.246
			Gemma-3-12B			
$< \lambda_a$	$< \lambda_b$	0.736	0.929	0.818	0.193	0.083
$< \lambda_a$	$\geq \lambda_b$	0.939	1.000	0.996	0.061	0.057
$\geq \lambda_a$	$< \lambda_b$	0.531	0.927	0.794	0.396	0.263
$\geq \lambda_a$	$\geq \lambda_b$	0.701	0.999	0.913	0.298	0.212
			Llama-3.1-8B			
$< \lambda_a$	$< \lambda_b$	0.786	0.928	0.709	0.141	-0.078
$< \lambda_a$	$\geq \lambda_b$	0.893	0.992	0.954	0.100	0.061
$\geq \lambda_a$	$< \lambda_b$	0.532	0.908	0.649	0.376	0.117
$\geq \lambda_a$	$\geq \lambda_b$	–	–	–	–	–

RC reduces the overconfidence of Direct Mean in the high-A_R, low-C_{RC} group on all three models. Nevertheless, the non-empty high-A_R groups and the low-A_R, high-C_{RC} groups retain positive RC gaps. Llama's low-A_R, low-C_{RC} group instead shifts from a Direct gap of 0.141 to an RC gap of -0.078: its absolute bias decreases, but the group becomes underconfident. Overall, RC effectively mitigates the overconfidence of Direct Mean, while the remaining positive gaps are concentrated in groups covered by the two RC-A trigger conditions. This supports applying a further adaptive correction after RC.

D.3 Response-Specific Rubric Information

Figure 13 reports all five metrics for the four estimators in Section 5.1 across all three models. All estimators use the same 3,910 candidate solutions per model. Matched Scaling and Shuffled Pairing follow the definitions in Section 5.1; the latter reports means over 2,000 permutations.

Across all three models, RC achieves the lowest Brier and the highest AUROC, PR-N, and PR-P among the four estimators; it also attains the lowest ECE on Qwen3-8B and Gemma-3-12B. Figure 4 visualizes the corresponding Qwen3-8B confidence distributions and reliability diagrams. These results show that response-specific pairing of direct confidence and rubric support provides benefits beyond uniform reduction or rubric reassignment.

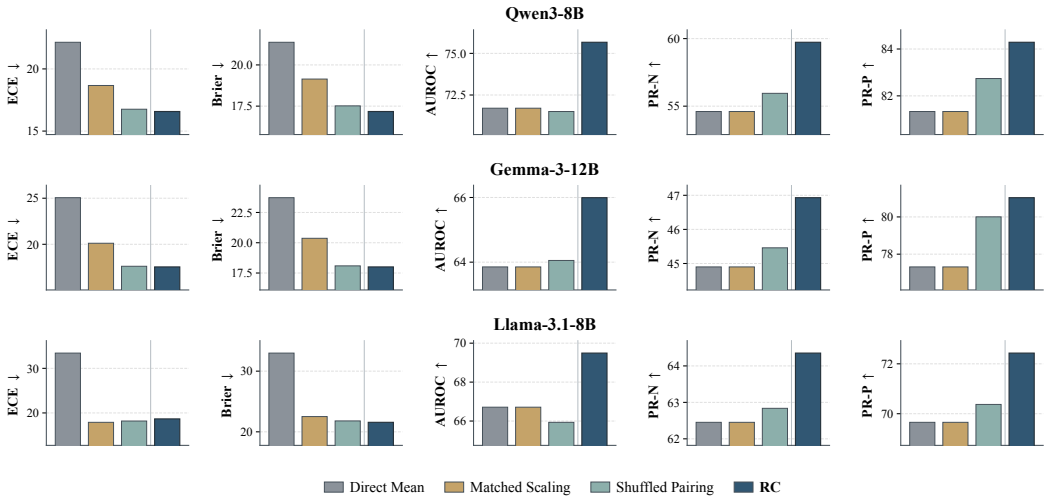


Figure 13: Response-specific rubric information across eight benchmarks (macro averages; all metrics are scaled by 10^2). Rows correspond to models; columns show ECE, Brier, AUROC, PR-N, and PR-P. Bars follow the legend order. Shuffled Pairing reports mean metrics over 2,000 permutations.