Analyzing Uncertainty of LLM-as-a-Judge: Interval Evaluations with Conformal Prediction

Anonymous EMNLP submission

Abstract

LLM-as-a-judge has become a promising paradigm for evaluating natural language generation (NLG), but the lack of reliability limits its deployment in high-risk applications. It has been very common to use LLMs to directly evaluate LLM-generated content while uncertainty quantification for rating evaluation remains underexplored. This work presents the first analysis framework to offer interval evaluations in LLM-based scoring via conformal prediction. Conformal prediction constructs continuous confidence intervals from a single evaluation run and we design a ordinal boundary adjustment for discrete rating tasks. We also suggest a midpoint-based score within the interval as a low-bias alternative to raw model score and weighted average. Extensive experiments and analysis across evaluators and conformal prediction methods show that our framework yields narrow intervals with reliable coverage, enabling more trustworthy evaluation for downstream decision making[1]

1 Introduction

Large language models (LLMs) have become powerful automatic evaluators for natural language generation (NLG) tasks, known as LLM-as-a-judge. Its consistency with human judgments results in strong performance with respect to metrics like ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2020). Besides, LLM judges can flexibly adapt to diverse evaluation criteria and provide scalable, cost-effective assessments compared to expert annotation (Gao et al., 2024; Gu et al., 2025). These advantages make the LLM-as-a-judge useful in various scenarios, such as clinical radiology (Chaves et al., 2024), rumor detection (Hong et al., 2025), cyberattack detection (Yong et al., 2025) and wildlife trafficking identification (Barbosa et al., 2025).

However, a single evaluation from a LLM judge might be biased (Wu and Aji, 2023; Li et al., 2024b) and uncertain due to inherent randomness (Schroeder and Wood-Doughty, 2024), thus undermining its reliability in scenarios like healthcare (Chung et al., 2025) and finance (Kamble et al., 2025). Though a LLM judge can express its confidence with well-designed prompt or via finetuning (Xu et al., 2024a; Liu et al., 2024; Taubenfeld et al., 2025), it may still suffer from overconfidence (Xiong et al., 2024) or dishonesty (Li et al., 2024d). We ask: *How can a LLM judge provide reliable evaluation given the user request?*

Conformal prediction (Vovk et al., 2005) is a promising way to quantify the uncertainty of an LLM judge (Ye et al., 2024). It outputs a confidence interval (or set for classification) to a model output with three key advantages. First, conformal prediction is a distribution-free uncertainty quantification method, which is suitable for black-box models like LLMs due to unknown input data distribution for most (if not all) LLMs. Second, it can provide post-hoc uncertainty quantification using only a calibration step based on LLM outputs. Third, the confidence interval given by conformal prediction enjoys statistically guaranteed coverage, i.e., how likely the ground truth falls within the interval, as long as the data is exchangeable.

In this paper, we comprehensively evaluate nine[2] conformal prediction methods in quantifying the uncertainty of a LLM judge in rating-based evaluation tasks, each of which constructs a confidence interval for a rating output by the LLM judge. For each conformal prediction method, we evaluate its efficiency (i.e., average width of confidence intervals) and coverage (i.e., the probability that ground truths fall within confidence intervals). Furthermore, to adapt to the ordinal and discrete nature in organic rating-based evaluation, we propose boundary adjustment that adjusts the endpoints of confi-

---

[1] Our code and data will be publicly available upon acceptance of this paper.

[2] Seven regression-type and two ordinal-type methods.

dence interval to be aligned with the rating scales. We prove that the boundary adjustment yields an interval suitable to the ordinal setting with provable non-decreasing coverage. From our comprehensive analysis, we demonstrate that the quality of confidence interval attributes to design choices of the LLM judge (e.g., which LLM to use as the judge, which prompting strategy for evaluation) as well as the size of calibration data during calibration. Finally, we show that the midpoint of the confidence interval provides better estimate to the ground truth to further assist better decision-making. Our analysis advocate for a shift from direct scoring to uncertainty-aware evaluations, offering references to decision-making.

In summary, our contributes are

- We are the first to analyze the uncertainty of LLM-as-a-judge using conformal prediction in rating-based evaluation, which uses the output from a single evaluation run.
- We design a boundary adjustment, which improves the efficiency empirically without compromising the coverage. The interval points suggest better alignment to human evaluation.
- We analyze factors affecting the interval quality, including the LLM-as-a-judge framework itself, the choice of LLM in the framework, and the size of calibration in conformal prediction, and offer practical insights or recommended choices.

2 Related Work

Uncertainty Quantification for LLM-as-a-Judge. Uncertainty quantification for LLM-as-a-judge is an important yet less explored area. Wagner et al. (2024) prompt the judge to justify each rating option as if it were correct and then construct a confusion matrix from token-level probabilities of these assessments to derive confidence scores. Xie et al. (2025) use token probabilities to estimate the confidence of judgments, and demonstrate that such measures exhibit bias and instability through extensive experiments. Similar conclusions are also found when applying other two common paradigms: (1) prompting LLMs to self-report confidence (Yona et al., 2024; Xu et al., 2024a), which can suffer from overconfidence (Xiong et al., 2024) or dishonesty (Li et al., 2024d); and (2) consistency-based approaches that rely on multiple generations (Tian et al., 2023; Xiong et al., 2024), which, like the confusion matrix-based method, are computationally expensive. To our best knowledge, Jung et al. (2024) is the most relevant work to us, which applies conformalized risk control (Angelopoulos et al., 2022) to ensure agreement with human preferences in pairwise response comparison (Zhou et al., 2024; Li et al., 2024a,c; Zhang et al., 2025; van den Burg et al., 2025). In contrast, we focus on using conformal prediction to quantify uncertainty in rating tasks instead of pairwise preference modeling.

Conformal Prediction for LLMs. Conformal prediction (Vovk et al., 2005) has drawn interest for uncertainty quantification in LLMs (Ye et al., 2024) due to its distribution-free and post-hoc nature with provable statistical guarantee. Owing to these advantages, recent works primarily apply conformal prediction to classification tasks, such as multiple-choice question answering (Kumar et al., 2023; Zhang et al., 2024; Su et al., 2024; Vishwakarma et al., 2025) and response selection for factual consistency (Quach et al., 2024; Mohri and Hashimoto, 2024; Wang et al., 2024; Kladny et al., 2025). These studies typically focus on ensuring that the correct answer is included in a unordered prediction set. However, we focus on providing intervals that reflect the variability in LLM judgements in rating tasks, which has ordinal preference.

3 Analyzing Uncertainty of LLM Judges

3.1 Preliminaries

LLM-as-a-Judge. In recent years, LLMs have been widely adopted as evaluators to score NLG tasks, which commonly yields a predicted score $\hat{y}$ on a Likert scale (Nemoto and Beglar, 2014). Following G-Eval (Liu et al., 2023), given a prompt p and a generated text x to be evaluated, an LLM judge M is expected to produce a response

$$M(p, x) = (z, \hat{y}), \tag{1}$$

where z denotes the logits $\hat{y}$ is a scalar score according to a predefined scale. Note that, for the logits, we only need to extract the logits of certain tokens (e.g., 1, 2, 3, 4, 5 if in a Likert scale) at the position of rating token only. Other than rating tasks, LLM-as-a-judge can also be applied to other evaluation paradigms (Gu et al., 2025), such as pairwise comparison or ranking, in which candidate outputs are first scored by the LLM judge and then compared or ordered based on those scores (Wang et al., 2025; Wei et al., 2025).

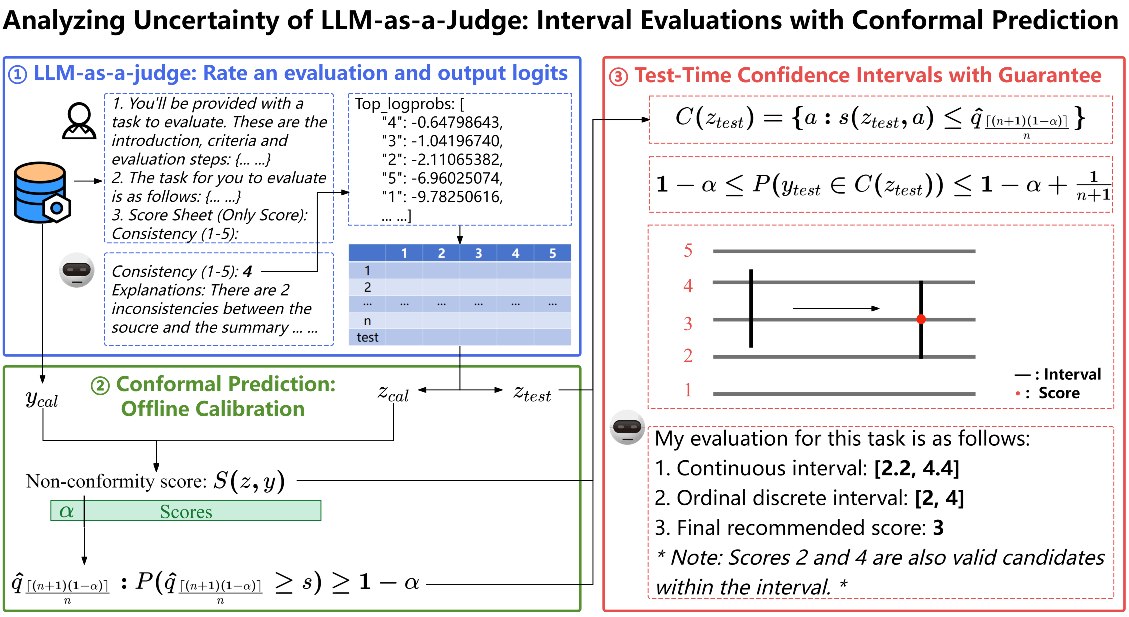
Analyzing Uncertainty of LLM-as-a-Judge: Interval Evaluations with Conformal Prediction
① LLM-as-a-judge: Rate an evaluation and output logits
1. You'll be provided with a task to evaluate. These are the introduction, criteria and evaluation steps: {... ...}
2. The task for you to evaluate is as follows: {... ...}
3. Score Sheet (Only Score): Consistency (1-5):
Top_logprobs: [
"4": -0.64798643,
"3": -1.04196740,
"2": -2.11065382,
"5": -6.96025074,
"1": -9.78250616,
... ...]
Consistency (1-5): 4
Explanations: There are 2 inconsistencies between the soucre and the summary ... ...
1 2 3 4 5
1
2
...
n
test
② Conformal Prediction: Offline Calibration
y_{cal}
z_{cal}
z_{test}
Non-conformity score: $S(z,y)$
α
Scores
$\hat{q}_{\frac{\lceil (n+1)(1-\alpha)\rceil}{n}}: P(\hat{q}_{\frac{\lceil (n+1)(1-\alpha)\rceil}{n}} \geq s) \geq 1-\alpha$
③ Test-Time Confidence Intervals with Guarantee
$C(z_{test}) = \{a : s(z_{test}, a) \leq \hat{q}_{\frac{\lceil (n+1)(1-\alpha)\rceil}{n}}\}$
$1-\alpha \leq P(y_{test} \in C(z_{test})) \leq 1-\alpha+\frac{1}{n+1}$
5
4
3
2
1
—: Interval
•: Score
My evaluation for this task is as follows:
1. Continuous interval: [2.2, 4.4]
2. Ordinal discrete interval: [2, 4]
3. Final recommended score: 3
* Note: Scores 2 and 4 are also valid candidates within the interval. *

Figure 1: Overview of quantifying the uncertainty in rating tasks. We apply conformal prediction to construct the confidence interval and set the width of the confidence interval as the uncertainty.

Conformal Prediction. Conformal prediction (Vovk et al., 2005) is a model-agnostic uncertainty quantification method. It constructs a confidence interval (or a set for classification) with coverage guarantee, free of training or prompting the judge model or assumptions about the underlying data distribution. In our work, we adopt split conformal prediction (Vovk et al., 2005), which quantify the uncertainty with a held-out calibration set. needs a offline calibration to capture the range of uncertainty. A non-conformity score function $s(z, y)$ is computed for each point in the calibration set, to measure how "unusual" a prediction $\hat{y}$ is to a ground truth y. For regression tasks, the non-conformity score is often defined as

$$s(z, y) = |\hat{y} - y|. \tag{2}$$

Given a user-desired miscoverage rate α, the $\frac{\lceil (n+1)(1-\alpha)\rceil}{n}$-quantile $\hat{q}$ of these scores is then used to construct the confidence interval for the prediction $\hat{y}_{test}$ of any test point

$$\mathcal{C}(z_{test}, \hat{y}_{test}) = [\hat{y}_{test} - \hat{q},\ \hat{y}_{test} + \hat{q}], \tag{3}$$

or equivalently

$$\mathcal{C}(z_{test}, \hat{y}_{test}) = \{a : s(z_{test}, a) \leq \hat{q}\}. \tag{4}$$

Such a confidence interval satisfies the coverage guarantee (Angelopoulos and Bates, 2022)

$$1-\alpha \leq \mathbb{P}\left(y_{test} \in \mathcal{C}\left(z_{test}, \hat{y}_{test}\right)\right) \leq 1-\alpha+\frac{1}{n+1}, \tag{5}$$

as long as the calibration set and test set are exchangeable, i.e., the joint distribution remain the same after any permutations on these two sets.

3.2 From Logits to Intervals

We focus on quantifying the uncertainty using conformal prediction in rating tasks (e.g., in Likert scale). An overview of the workflow is presented in Figure 1.

Extract Logits as Feature As our framework targets uncertainty estimation in discrete rating tasks, the token-level logits corresponding to Likert-scale scores (e.g., 1–5) are used as features.

As shown in Figure 1, an LLM judge, prompted with a chain-of-thought (CoT) instruction (Example in Appendix A.1) that specifies an output format, generates a response containing its rating. After accurately locating the target score token "4", we extract the log probabilities of all potential score tokens (e.g., 1–5). To ensure semantic consistency, we aggregate the log probabilities of tokens with equivalent meanings.

As a result, we obtain a K-dimensional feature vector z representing the logits associated with each candidate score token in $\{1, 2, \ldots, K\}$[3], which composes the input for conformal prediction,

---

[3] We use $K = 5$ for the standard Likert scale or GPA-like settings, but K can be adapted to other granularities (e.g., 7 or 10) depending on the evaluation scale.

i.e. $\{z, y\}$, which is found to have these properties: (1) Independent and identically distributed, at least exchangeable[4]; (2) $\mathbb{E}[\Sigma_{i=1}^{k} e^{z_i}] = 1$, which causes the interdependence among variables; (3) heteroskedasticity[5]; (4) Isolated distribution of label caused by rating nature. These properties inspires our method choice for interval construction.

Notably, token probabilities are frequently seen in early works but they would cause multicollinearity in regression. Thus they are only used in ordinal predictors (because they need) and cause unstable performance due to well-known bias.

Confidence Interval Estimation As discussed in Section 3.1, modern conformal prediction methods vary in how they define non-conformity scores and construct intervals, yet they share a unified structure that ensures valid coverage. In our framework, we go beyond the basic absolute-error formulation and adopt a diverse set of nine conformal predictors. Each is designed to handle specific data characteristics—such as asymmetry, heteroskedasticity, or ordinal outputs—and offers complementary strengths. This diversity allows our framework to remain robust and adaptable across different evaluation scenarios.

The selected methods include quantile regression–based approaches (e.g., CQR (Romano et al., 2019), Asymmetric CQR (Sesia and Candès, 2019)), histogram-based estimators (e.g., CHR (Sesia and Romano, 2021)), kernel regression variants (e.g., LVD (Lin et al., 2021)), boosted methods (Xie et al., 2024) (e.g., Boosted CQR and Boosted LCP), and ordinal classification–based predictors (e.g., R2CCP (Guha et al., 2024), Ordinal APS (Lu et al., 2022) and Ordinal Risk Control (Xu et al., 2024b)). A complete summary of their non-comformity score functions and construction principles is provided in Appendix A.4.

3.3 Boundary Adjustment

In addition to an interval in Equation 3, our framework further transform the regression problem to a ordinal classification problem by a boundary adjustment, due to the ordinal and discrete nature of rating. Therefore the interval boundaries will be aligned with potential labels, instead of continuous scores that might have no exact meaning.

For conformal prediction, we redefine the non-conformity score function as:

$$s'(z,y) = s(z,y') = \begin{cases} s(z,\lceil y \rceil) & \text{if } y \le \lfloor \hat{y} \rfloor, \\ s(z,y) & \text{otherwise,} \\ s(z,\lfloor y \rfloor) & \text{if } y \ge \lceil \hat{y} \rceil. \end{cases} \tag{6}$$

Because all potential labels y' are integers in rating evaluation, this new function ensures the scores consistent on calibration set. However, it transforms the interval from Equation 4 to

$$\mathcal{C}(z_{test}) = \{a : s'(x_{test}, a) \le \hat{q}\} = [l,u] \to [l',u'], \tag{7}$$

where $l' = \lceil l \rceil$ and $u' = \lfloor u \rfloor$.

We shrink the boundaries to interior labels by cutting excessive areas because they cover no potential labels, while y in these areas share a same $\hat{q}$ with the labels, which means this adjustment has no influence to coverage. For example, $[2.9, 4.2]$ will be shrunk to $[3, 4]$ with the same coverage since $s'(z, 2.9) = s'(z, 3)$ and $s'(z, 4.2) = s'(z, 4)$.

On the other hand, we can also expand an interval to mitigate the marginal miscoverage caused by isolated label distribution and limited calibration size. For example, the interval $[2.2, 3.9]$ only cover 3 but can be expanded to $[2, 4]$, then a pitiful miscoverage can be avoided if the ground truth is 2 or 4. This improvement of coverage can be explained by a larger $\hat{q}$ for each boundary, which theoretically ensures more abnormal results to be covered.

The following theorem shows the non-decreasing coverage after boundary adjustment, with its proof shown in Appendix A.2. Other discrete ganularities (e.g. GPA scale) are also applicable after linear transformation to intergers.

Theorem 1 (Non-decreasing Coverage After Boundary Adjustment)**.** *Based on coverage guarantee in Equation 5, we transform the non-conformity score function $s(x, y)$ by Equation 6 and adjust an continuous interval by Equation 7.*

Then, if the adjustment is performed by shrinking ($l' = \lceil l \rceil$ and $u' = \lfloor u \rfloor$), coverage preserves:

$$\mathbb{P}\Big(Y_{test} \in \mathcal{C}'(x_{test})\Big) \ge 1 - \alpha.$$

And if at least one boundary is expanded ($l' = \lfloor l \rfloor$ or $u' = \lceil u \rceil$), coverage is expected to increase:

$$\mathbb{P}\Big(Y_{test} \in \mathcal{C}'(x_{test})\Big) > 1 - \alpha.$$

---

[4] It depends on the generation task to evaluate. If the dateset comprises summaries and annotations generated by multiple models from the same source document, the features extracted by the LLM judge are unlikely to be i.i.d. However, exchangeability is guaranteed for permutation invariance of evaluations.

[5] Hypothesis testing results are shown in Appendix A.3.

3.4 Midpoints as Calibrated Scores

To make use of confidence interval, one can take its midpoint as a suggested score. The interval provides coverage guarantees but offers no indication of the direction toward the true label. Even if biased, the midpoint is the minimum-variance estimator of the true label given the endpoints.

4 Experiments

4.1 Experimental Setting

Datasets. We run experiments on evaluation benchmarks in text summarization, dialogue summarization and reasoning. For summarization, we use SummEval (Fabbri et al., 2021) (1,600 samples) and DialSumm (Gao and Wan, 2022) (1,400 samples), each annotated by three human raters using Likert-scale scores across four dimensions. The average of the three ratings is used as the ground-truth label on a GPA scale. For reasoning, we use the annotations of overall quality for CosmosQA (Li et al., 2023), DROP (Dua et al., 2019), e-SNLI (Camburu et al., 2018) and GSM8K (Cobbe et al., 2021) in ROSCOE (Golovneva et al., 2023), each with around 200 samples on Likert scale.

LLM-based Evaluation. We primarily adopt G-Eval (Liu et al., 2023) as our judge framework with a CoT prompt (Appendix A.1). For reasoning tasks, we additionally employ SocREval (He et al., 2024). Evaluations are mainly conducted using GPT-4o-mini (2024-07-18), DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025), and Qwen2.5-72B-Instruct (Qwen et al., 2025), all of which provide token-level logits.

Conformal Prediction. In our experiments, we employ 9 regression type conformal predictors introduced in Section 3.2. Detailed introduction of each method is provided in Appendix A.4. For each method, we split of dataset into 50% calibration set and 50% test set and run it 30 times with different random seeds and report the mean and standard deviation of interval width and coverage rate.

Human-based Baseline. Due to lack of reference, we design a human-based baseline by Equation (2) and (3) after randomly choosing one annotation as prediction. Comparison on two datasets are discussed in Appendix A.5.

4.2 Continuous Intervals Indicate Uncertainty

Across all experimental settings (Table 1 and Table 7), most conformal predictors consistently generate confidence intervals with coverage close to the 90%, indicating LLM judges were more certain when evaluating summarization tasks, especially on the ’Fluency’ of the summaries. However, some methods show unsatisfying coverage especially in small-sample conditions (ROSCOE), further highlighting the importance of proper calibration (see Section 5.1). To mitigate miscoverage, a mild adjustment (e.g. adjust to nereast label within 0.1 radius) is effective (Table 9 and 10).

4.3 All Coverages Improve after Adjustment

While continuous interval estimators offer intuitive uncertainty representations, their coverage is highly sensitive to the variability of quantile estimation in calibration. Table 2 and 8 show that this issue is well addressed by boundary adjustment: all conformal regressors exhibit increased coverage after adjustment, with the majority achieving stable coverage at or above the 90% across datasets and evaluation frameworks.

With theoretical guarantee (Theorem 1), coverage improvements brought by boundary adjustment are empirically robust across different datasets, conformal predictors, judge models and judgment frameworks. For example, in SummEval and DialSumm, where coverage originally fell in the 83%–88% range, predictors now consistently exceed 90%. A significant example is LVD, which, when applied to e-SNLI using Qwen2.5 under SocREval, showed an increase from 85.96% to 95.53%, demonstrating the practical effectiveness.

4.4 Recommended Choice from Analysis

Our analysis shows that DeepSeek-R1-Distill-Qwen-32B provides the most consistent coverage (surpassing Qwen2.5-72B-Instruct and GPT-4o-mini), while Qwen2.5-72B-Instruct typically yields the narrowest intervals; under the G-Eval framework we observe higher coverage (at the cost of slightly wider bands) versus SocREval; among CP methods, R2CCP strikes the best balance between coverage and width, Boosted LCP performs comparably but less efficiently, LVD delivers very tight intervals without sacrificing coverage, and OrdinalAPS can produce the tightest bands yet is highly sensitive to token probabilities. In practice, for high-risk applications we recommend DeepSeek +

Evaluator	Method	SummEval Evaluated with G-Eval				ROSCOE Evaluated with SocREval			
		Consistency	Coherence	Fluency	Relevance	CosmosQA	DROP	e-SNLI	GSM8K
GPT-4o mini	Boosted CQR	1.01 / 87.75%	2.73 / 87.80%	1.54 / 88.68%	2.00 / 87.42%	3.15 / 80.07%	2.63 / 78.57%	1.82 / 80.26%	3.08 / 82.50%
	Boosted LCP	0.76 / 89.22%	2.67 / 87.34%	0.92 / 89.18%	1.91 / 87.19%	3.60 / 83.91%	2.92 / 85.40%	1.88 / 81.23%	3.36 / 85.93%
	CHR	0.67 / 88.99%	2.41 / 82.96%	0.94 / 88.86%	1.74 / 82.62%	2.54 / 73.06%	1.86 / 68.92%	1.36 / 72.24%	1.98 / 78.67%
	Asym CQR	1.25 / 94.97%	2.91 / 93.76%	1.60 / 93.75%	2.13 / 91.42%	3.90 / 98.71%	3.91 / 98.60%	**2.87/ 96.67%**	3.89 / 98.80%
	Sym CQR	1.15 / 94.16%	**2.87 / 93.15%**	1.44 / 92.92%	**2.09 / 90.92%**	**3.53 / 95.27%**	**3.82/96.70%**	3.04 / 96.62%	**3.53 / 95.67%**
	LVD	1.01 / 92.35%	2.73 / 89.76%	**1.11 / 90.59%**	2.02 / 89.55%	3.10 / 83.95%	2.49 / 83.05%	2.17 / 86.18%	3.08 / 89.57%
	R2CCP	**0.69 / 90.88%**	2.62 / 89.63%	0.92 / 89.36%	1.97 / 89.70%	2.96 / 85.85%	2.43 / 84.73%	1.75 / 84.02%	2.15 / 85.07%
DSR1-Distill-Qwen-32B	Boosted CQR	1.10 / 89.30%	2.36 / 88.98%	1.16 / 89.46%	2.00 / 88.98%	3.17 / 82.72%	2.47 / 81.11%	1.79 / 80.96%	2.94 / 79.83%
	Boosted LCP	0.77 / 89.20%	2.32 / 86.70%	0.93 / 89.10%	1.91 / 86.89%	3.48 / 81.60%	2.79 / 85.46%	1.84 / 80.61%	3.43 / 85.23%
	CHR	0.82 / 91.17%	2.23 / 87.07%	0.90 / 89.24%	1.87 / 86.38%	2.66 / 76.50%	1.95 / 78.06%	1.38 / 71.97%	2.01 / 81.60%
	Asym CQR	1.30 / 95.13%	2.72 / 92.86%	1.49 / 94.52%	2.21 / 92.06%	3.84 / 99.08%	3.95 / 99.27%	2.86 / 96.05%	3.85 / 98.43%
	Sym CQR	1.16 / 93.88%	2.67 / 92.50%	1.31 / 93.01%	2.13 / 91.05%	**3.48 / 96.70%**	**3.83 / 96.35%**	2.97 / 96.36%	3.46 / 95.60%
	LVD	0.97 / 92.93%	2.43 / 91.10%	1.00 / 91.10%	**2.04 / 90.14%**	3.25 / 88.10%	2.62 / 88.06%	**2.24 / 90.96%**	**3.02/90.63%**
	R2CCP	**0.69 / 90.44%**	**2.30 / 90.12%**	**0.89 / 90.09%**	2.00 / 89.84%	2.94 / 86.97%	2.29 / 86.35%	1.85 / 87.87%	1.88 / 85.33%
Qwen2.5-72B-Instruct	Boosted CQR	0.80 / 88.28%	2.46 / 87.82%	1.24 / 89.22%	1.88 / 87.17%	3.05 / 79.08%	2.56 / 81.17%	1.51 / 77.11%	2.81 / 80.67%
	Boosted LCP	0.67 / 88.81%	2.43 / 86.92%	0.94 / 89.26%	1.86 / 87.51%	3.46 / 80.41%	2.81 / 85.75%	1.74 / 77.50%	3.38 / 86.23%
	CHR	0.61 / 89.04%	2.14 / 80.93%	0.98 / 88.93%	1.61 / 79.61%	2.44 / 72.65%	2.08 / 75.87%	1.22 / 69.69%	1.81 / 77.50%
	Asym CQR	1.11 / 94.47%	2.80 / 93.13%	1.63 / 94.79%	2.17 / 92.21%	3.86 / 99.01%	3.89 / 98.67%	**2.77 / 96.84%**	3.87 / 98.97%
	Sym CQR	0.98 / 93.10%	2.73 / 92.25%	1.44 / 93.73%	2.11 / 91.30%	**3.37 / 94.80%**	3.79 / 97.02%	3.01 / 97.37%	3.35 / 95.33%
	LVD	0.85 / 92.82%	2.55 / 90.49%	1.09 / 90.94%	1.94 / 89.27%	3.05 / 84.29%	**2.67 / 90.57%**	1.91 / 85.96%	**2.83/ 90.13%**
	R2CCP	**0.61 / 90.73%**	2.44 / 89.54%	**0.95 / 90.17%**	**1.98 / 90.45%**	2.90 / 85.34%	2.39 / 86.25%	1.59 / 84.50%	2.00 / 86.73%

Table 1: Summaries of continuous intervals, SummEval evaluated by G-Eval and ROSCOE evaluated by SocREval.

G-Eval + LVD, whereas, when employing boundary adjustment to reinforce coverage, the most efficient setup is Qwen2.5 + R2CCP + SocREval.

4.5 Midpoints Reduce at Most 95% Bias

A simple yet effective way to leverage these interval estimates is to use their midpoints as point predictions, assuming that the true score lies near the center of a sufficiently narrow interval, with extreme cases yielding exact overlap. In the interval estimation phase, R2CCP is found to be the best predictor with proper calibration, although it was occasionally outperformed by OrdinalAPS. Based on this, we select three interval types for midpoint evaluation: the continuous and discrete (adjusted) intervals of R2CCP, and the discrete interval derived directly from OrdinalAPS. We compare these midpoint scores against two baselines: the raw score in LLM response and the weighted average derived from token probabilities.

Table 3 and 11 show that, while the midpoint estimates achieve comparable or even slightly better correlation with the ground-truth scores, they significantly reduce prediction error, yielding substantially lower MSE and MAE. For example, the midpoints from OrdinalAPS on Coherence of DialSumm evaluated by DeepSeek-R1-Distill-Qwen-32B reduce 95.7% of the MSE, from 2.908 to 0.124 (Table 11). Moreover, across multiple dimensions, the MAE between midpoints and the true ratings is typically below 0.5 (in some cases drops to around 0.2), demonstrating excellent debiasing. For instance, when the ground truth is 4.33, the midpoint could be 4.0 or 4.5. This suggests that midpoints derived from narrow intervals offer a more accurate and robust alternative for score estimation directly from rating-type LLM-as-a-judge.

5 Analysis

5.1 Proper Calibration Improves Coverage

Due to the limited sample size in ROSCOE and the inability of most CP methods to achieve the target coverage, we turn to the SummEval and DialSumm datasets to examine how calibration-set size affects coverage. Using R2CCP, we construct continuous prediction intervals under four calibration regimes—25%, 50%, 75%, and 100% of the whole calibration set, and show the tendency in Figure 2, which show that as the calibration set grows, the mean coverage increases and even surpasses the desired 90%, while its corresponding error bar shrinks. This behavior highlights the importance of sufficiently large calibration sets to stabilize coverage around the required threshold.

5.2 Why Boundary Adjustment is Effective

During the transition from continuous to discrete intervals, we observe consistent improvements in empirical coverage across all experimental settings. A concrete example is illustrated in Figure 3 and 4, where certain ground-truth scores fall just outside the estimated intervals. In such cases, a marginal upward shift in the estimated quantiles would suffice to restore coverage. This demonstrates that a relatively modest increase in interval width can lead to a substantial gain in calibration, successfully achieving the 90% coverage.

Evaluator	Method	SummEval Evaluated with G-Eval				ROSCOE Evaluated with SocREval			
		Consistency	**Coherence**	**Fluency**	**Relevance**	**CosmosQA**	**DROP**	**e-SNLI**	**GSM8K**
GPT-4o mini	Boosted CQR	0.99 / 92.81%	2.73 / 93.02%	1.54 / 94.38%	2.00 / 92.93%	3.20 / 93.40%	2.63 / 89.65%	1.82 / 92.15%	3.09 / 91.17%
	Boosted LCP	0.74 / 91.90%	2.68 / 93.53%	**0.90 / 90.88%**	**1.91 / 92.70%**	3.60 / 95.48%	3.01 / 91.27%	1.90 / 91.80%	3.26 / 92.17%
	CHR	0.70 / 91.79%	2.41 / 87.78%	0.94 / 90.60%	1.74 / 88.10%	2.56 / 82.45%	1.87 / 78.86%	1.34 / 83.46%	1.94 / 83.23%
	Asym CQR	1.25 / 96.02%	2.90 / 95.41%	1.60 / 94.57%	2.14 / 94.14%	3.90 / 98.84%	3.91 / 98.73%	2.87 / 96.89%	3.89 / 98.80%
	Sym CQR	1.15 / 95.45%	2.87 / 94.94%	1.44 / 93.80%	2.09 / 93.56%	3.53 / 95.34%	3.82 / 97.05%	3.04 / 96.89%	3.53 / 95.67%
	LVD	1.01 / 94.11%	2.73 / 93.72%	1.12 / 92.70%	2.03 / 93.82%	**3.13 / 91.53%**	**2.52 / 90.22%**	2.17 / 94.82%	**3.09/93.37%**
	R2CCP	**0.68 / 92.15%**	**2.62 / 92.81%**	0.91 / 90.99%	1.97 / 93.38%	2.93 / 89.46%	2.41 / 89.21%	**1.71 / 90.11%**	2.09 / 86.93%
	OrdinalAPS	2.28 / 42.60%	1.88 / 70.90%	1.78 / 71.47%	2.36 / 51.85%	0.73 / 47.52%	0.83 / 55.08%	0.72 / 52.76%	0.58 / 73.90%
	OrdinalRC	2.41 / 75.19%	2.02 / 67.38%	1.93 / 14.58%	2.51 / 90.30%	0.82 / 49.46%	0.91 / 57.11%	0.80 / 54.61%	0.60 / 74.43%
DSR1-Distill-Qwen-32B	Boosted CQR	1.08 / 93.55%	2.37 / 93.96%	1.15 / 93.48%	2.01 / 93.72%	3.20 / 95.71%	**2.52 / 93.30%**	**1.79 / 93.25%**	2.94 / 92.23%
	Boosted LCP	0.76 / 92.03%	2.32 / 92.37%	0.93 / 91.34%	1.92 / 92.81%	3.46 / 95.95%	2.80 / 91.94%	1.87 / 92.89%	3.36 / 93.63%
	CHR	0.87 / 93.96%	2.23 / 91.42%	0.91 / 91.98%	1.87 / 90.84%	2.69 / 86.80%	1.97 / 85.90%	1.39 / 85.96%	2.01 / 86.60%
	Asym CQR	1.31 / 95.99%	2.72 / 94.83%	1.49 / 95.57%	2.21 / 94.53%	3.84 / 99.08%	3.95 / 99.27%	2.88 / 96.45%	3.84 / 98.47%
	Sym CQR	1.15 / 95.02%	2.67 / 94.34%	1.32 / 94.44%	2.13 / 93.67%	3.48 / 96.80%	3.82 / 96.54%	2.99 / 96.80%	3.46 / 95.63%
	LVD	0.97 / 95.01%	2.44 / 94.58%	1.00 / 93.21%	2.04 / 94.12%	3.28 / 95.27%	2.67 / 93.75%	2.24 / 96.36%	3.03 / 94.40%
	R2CCP	0.68 / 91.57%	2.30 / 93.22%	**0.89 / 91.80%**	1.99 / 92.96%	**2.91 / 90.58%**	2.25 / 89.97%	**1.80 / 92.35%**	1.82 / 86.93%
	OrdinalAPS	**0.61 / 89.84%**	**0.66 / 90.70%**	1.14 / 90.75%	**0.54 / 90.11%**	1.32 / 60.00%	1.26 / 78.22%	1.46 / 87.85%	1.50 / 85.67%
	OrdinalRC	2.54 / 90.11%	2.56 / 91.18%	3.73 / 89.53%	2.14 / 90.07%	1.44 / 62.35%	1.33 / 78.22%	1.52 / 88.33%	1.55 / 86.07%
Qwen2.5-72B-Instruct	Boosted CQR	0.81 / 92.36%	2.47 / 93.06%	1.25 / 93.66%	1.88 / 92.81%	3.10 / 94.01%	2.56 / 90.79%	**1.49 / 92.11%**	2.82 / 92.03%
	Boosted LCP	0.65 / 91.26%	2.44 / 92.26%	0.93 / 91.20%	1.86 / 92.57%	3.40 / 94.90%	2.84 / 92.41%	1.79 / 91.84%	3.33 / 92.90%
	CHR	0.66 / 92.21%	2.14 / 86.10%	0.98 / 91.16%	1.61 / 85.78%	2.49 / 82.14%	2.05 / 82.89%	1.18 / 84.56%	1.79 / 85.27%
	Asym CQR	1.10 / 95.47%	2.79 / 94.70%	1.64 / 95.63%	2.17 / 94.85%	3.85 / 99.18%	3.89 / 98.67%	2.77 / 97.06%	3.87 / 98.97%
	Sym CQR	0.98 / 94.35%	2.72 / 94.18%	1.45 / 94.79%	2.10 / 94.02%	3.36 / 95.07%	3.79 / 97.08%	3.01 / 97.68%	3.34 / 95.33%
	LVD	0.85 / 95.11%	2.56 / 94.05%	1.09 / 93.45%	1.95 / 93.86%	**3.07 / 92.01%**	2.67 / 93.87%	1.91 / 95.53%	2.87 / 93.43%
	R2CCP	**0.59 / 91.83%**	2.43 / 92.78%	**0.95 / 92.12%**	1.98 / 93.72%	2.88 / 89.29%	**2.34 / 90.00%**	1.55 / 90.20%	1.96 / 88.57%
	OrdinalAPS	0.80 / 89.94%	**0.77 / 90.60%**	0.90 / 58.30%	**0.62 / 90.40%**	0.71 / 55.99%	0.25 / 56.83%	0.67 / 77.68%	0.46 / 70.87%
	OrdinalRC	2.85 / 90.00%	2.96 / 89.35%	3.21 / 53.31%	2.75 / 90.14%	0.75 / 57.28%	0.29 / 56.83%	0.80 / 79.74%	0.49 / 71.37%

Table 2: With boundary adjustment, nearly all intervals generated by conformal regressors meet or exceed the 90% coverage on SummEval evaluated by G-Eval and ROSCOE evaluated by SocREval.

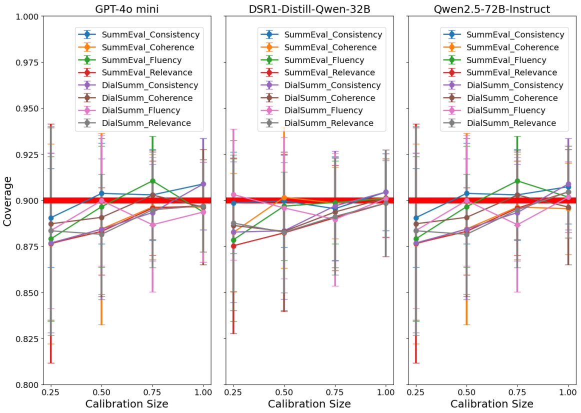

Figure 2: Coverage improves as calibration increases: mean coverage rates increase to 90% and error bars shrink on all dimensions of SummEval and DialSumm.

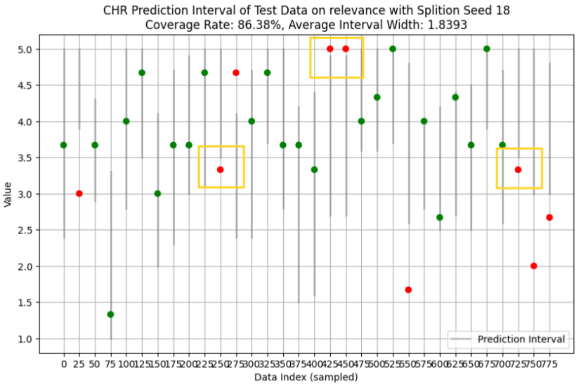

Figure 3: Red points mean the labels lying outside the intervals, which could turn green (inside) if the interval just extend to nearest labels (e.g. 3.33 and 5).

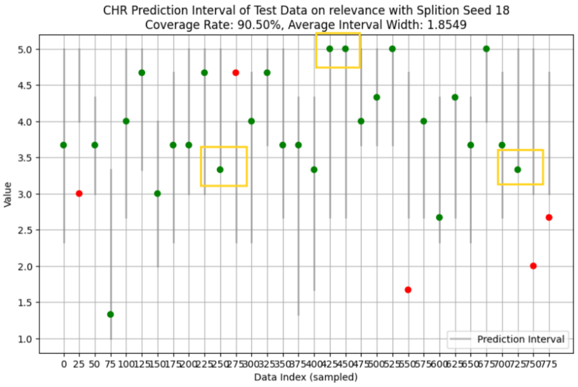

Figure 4: After applying boundary adjustment, the coverage in this instance improves from 86.38% to 90.50%, while the average width increases slightly to 1.8549.

5.3 Reprompt and Regrade with Intervals

Assuming LLM judge could mimic human to decide, we explore the potential of intervals in decision-making by reprompting judges with intervals information (Figure 6). We reprompt our best intervals among 30 experiments of ROSCOE (R2CCP + DSR1) to the judge and find that intervals strengthen its confidence in initial ratings , which mostly lies within the intervals (Table 14, Figure 7, 8 and 9). And it also retains its rating even it's outside the interval (Figure 10 and 11).

Evaluator	Method	Coherence				Consistency				Fluency				Relevance			
		MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ
GPT-4omini	Raw Score	1.729	1.055	0.446	0.373	1.674	1.073	0.480	0.437	7.099	2.393	0.051	0.045	1.009	0.786	0.512	0.427
	Weighted Sum	1.643	1.037	0.514	0.379	1.548	1.066	0.478	0.383	3.412	1.733	0.319	0.250	0.865	0.737	0.567	0.419
	OrdinalAPS	**1.004**	**0.895**	0.511	**0.412**	2.157	1.377	**0.481**	0.403	**0.872**	**0.792**	0.315	**0.265**	2.090	1.351	0.538	**0.441**
	Con_R2CCP	**0.791**	**0.716**	0.512	0.373	**0.510**	**0.432**	0.455	0.371	**0.442**	**0.491**	**0.330**	**0.261**	**0.418**	**0.509**	0.546	0.403
	Dis_R2CCP	**0.794**	**0.715**	0.508	**0.386**	**0.512**	**0.428**	**0.506**	**0.468**	**0.443**	**0.488**	**0.336**	**0.300**	**0.423**	**0.509**	0.540	0.423
DSR1-Qwen32	Raw Score	1.010	0.775	0.549	0.457	1.229	0.770	0.467	0.425	2.843	1.549	0.387	0.355	0.763	0.682	0.520	0.437
	Weighted Sum	0.869	0.734	0.599	0.447	1.439	1.065	0.468	0.375	2.783	1.564	0.420	0.332	0.646	0.632	0.565	0.419
	OrdinalAPS	**0.096**	**0.224**	0.570	**0.489**	**0.147**	**0.322**	**0.471**	0.424	**0.315**	**0.528**	0.377	0.345	**0.076**	**0.207**	0.534	**0.458**
	Con_R2CCP	**0.599**	**0.619**	**0.663**	**0.492**	**0.564**	**0.446**	0.445	0.361	**0.373**	**0.455**	0.391	0.311	**0.431**	**0.513**	0.555	0.412
	Dis_R2CCP	**0.602**	**0.619**	**0.661**	**0.508**	**0.566**	**0.441**	0.462	0.423	**0.375**	**0.454**	0.393	0.351	**0.434**	**0.512**	0.548	0.431
Qwen2.5-72B	Raw Score	1.432	0.981	0.426	0.358	2.068	1.237	0.458	0.416	4.476	1.958	0.310	0.281	1.188	0.903	0.498	0.420
	Weighted Sum	1.282	0.932	0.539	0.395	1.847	1.213	0.483	0.387	4.236	1.928	0.363	0.285	1.091	0.885	0.555	0.412
	OrdinalAPS	**0.111**	**0.265**	0.526	**0.449**	**0.201**	**0.408**	**0.497**	**0.453**	**0.388**	**0.574**	0.335	**0.297**	**0.083**	**0.215**	0.539	**0.470**
	Con_R2CCP	**0.675**	**0.659**	**0.603**	**0.444**	**0.469**	**0.396**	0.465	0.378	**0.414**	**0.486**	0.340	0.269	**0.407**	**0.502**	**0.571**	**0.425**
	Dis_R2CCP	**0.678**	**0.659**	**0.600**	**0.456**	**0.469**	**0.387**	**0.538**	**0.498**	**0.416**	**0.485**	0.342	**0.306**	**0.411**	**0.501**	**0.566**	**0.444**

Table 3: Comparison of interval midpoints with LLM scoring baselines on SummEval. **Bold** indicates better performance than baselines, underlined denotes comparable performance, and gray indicates worse performance. Mean squared error (MSE), mean absolute error (MAE), Spearman's ρ and Kendall's τ are calculated as the mean of 30 experiments. The R2CCP midpoints consistently yield significant improvements in MSE and MAE, though less dramatic than those of OrdinalAPS (e.g., MSE < 0.1 in relevance). Notably, when interval quality is poor, midpoint predictions also degrade, underscoring the importance of proper calibration (Table 12 and 13).

6 Discussion

This work introduces the application of Conformal Prediction (CP) methods for quantifying LLM scoring uncertainty based on single-output logits. We provide the first analysis of applying CP in estimating LLM scoring uncertainty: by employing nine distinct CP methods across three LLM judge models, two evaluation frameworks, and multiple datasets, we construct continuous and discrete prediction intervals that achieve or approximate 90% confidence coverage. Moreover, we design a theoretically grounded boundary adjustment technique that transforms continuous intervals to discrete rating scales, yielding a global improvement in coverage and enhancing the reliability and interpretability of the intervals. Finally, we explore using interval midpoints as calibrated scores to assess the utility of interval estimation within the LLM-as-a-judge paradigm. Experimental results demonstrate that this strategy matches or slightly surpasses baselines on correlation metrics while significantly outperforming direct scoring on error metrics, thereby achieving higher accuracy.

To make use of the intervals, the quantified uncertainty by confidence intervals is the key. As long as an LLM judge is used for evaluation, an interval evaluation helps users to determine when they can trust the judgment. On the one hand, a wider confidence interval serves as a warning signal of unreliability with the score, which is particularly beneficial in high-risk environments where uncertainty-induced errors must be minimized, such as in medical diagnosis (Lu et al., 2022; Tan et al., 2024). On the other hand, a narrower confidence interval suggests a higher degree of certainty in the score, thereby indicating selective prediction and reducing the need for manual review in automated evaluation, such as in essay scoring (Song et al., 2024). We believe our framework might be helpful in example selection to avoid the model collapse when trained on LLM generated data (Shumailov et al., 2024), since wide intervals could contribute to active learning.

Limitations

This paper has main limitations in tasks for LLM to judge. Our experimental results are primarily based on summarization and reasoning in NLG tasks, with a focus on the SummEval, DialSumm and ROSCOE. Additionally, we acknowledge that there are numerous other tasks that we have yet to explore, including but not limited to machine translation, multimodal generation, etc.

References

Anastasios N. Angelopoulos and Stephen Bates. 2022. A gentle introduction to conformal prediction and distribution-free uncertainty quantification.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2022. Learn then test: Calibrating predictive algorithms to achieve risk control.

Juliana Barbosa, Ulhas Gondhali, Gohar A. Petrossian, Kinshuk Sharma, Sunandan Chakraborty, Jennifer

Jacquet, and Juliana Freire. 2025. A cost-effective llm-based approach to identify wildlife trafficking in online marketplaces.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Hassan Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. 2024. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature communications*, 16 1:3108.

Philip Chung, Akshay Swaminathan, Alex J. Goodell, Yeasul Kim, S. Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, David Seong, Andrew A. Lee, Caitlin E. Coombes, Brad Bradshaw, Mahir A. Sufian, Hyo Jung Hong, Teresa P. Nguyen, Mohammad R. Rasouli, Komal Kamra, Mark A. Burbridge, James C. McAvoy, Roya Saffary, Stephen P. Ma, Dev Dash, James Xie, Ellen Y. Wang, Clifford A. Schmiesing, Nigam Shah, and Nima Aghaeepour. 2025. Verifact: Verifying facts in llm-generated clinical text with electronic health records.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

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

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan

Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Roscoe: A suite of metrics for scoring step-by-step reasoning.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Leying Guan. 2022. Localized conformal prediction: A generalized inference framework for conformal prediction.

Etash Kumar Guha, Shlok Natarajan, Thomas Möllenhoff, Mohammad Emtiyaz Khan, and Eugene Ndiaye. 2024. Conformal prediction via regression-as-classification. In *The Twelfth International Conference on Learning Representations*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2024. Socreval: Large language models with the socratic method for reference-free reasoning evaluation.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *THE WEB CONFERENCE 2025*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 8487–8495. Association for Computational Linguistics.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: Llm judges with provable guarantees for human agreement.

Kiran Kamble, Melisa Russak, Dmytro Mozolevskyi, Muayad Ali, Mateusz Russak, and Waseem AlShikh. 2025. Expect the unexpected: Failsafe long context qa for finance.

Klaus-Rudolf Kladny, Bernhard Schölkopf, and Michael Muehlebach. 2025. Conformal generative modeling with improved sample efficiency through sequential greedy filtering.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2024a. Calibraeval: Calibrating prediction distribution to mitigate selection bias in llms-as-judges.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024b. Llms-as-judges: A comprehensive survey on llm-based evaluation methods.

Minzhi Li, Zhengyuan Liu, Shumin Deng, Shafiq Joty, Nancy F. Chen, and Min-Yen Kan. 2024c. Dna-eval: Enhancing large language model evaluation through decomposition and aggregation.

Shiyang Li, Jianshu Chen, and Dian Yu. 2023. Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach.

Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, Jie Zhou, Yujiu Yang, Ngai Wong, Xixin Wu, and Wai Lam. 2024d. A survey on the honesty of large language models.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2021. Locally valid and discriminative prediction intervals for deep learning models. In *Advances in Neural Information Processing Systems*, volume 34, pages 8378–8391. Curran Associates, Inc.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2024. Can LLMs learn uncertainty on their own? expressing uncertainty effectively in a self-training manner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21635–21645, Miami, Florida, USA. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2303.16634*.

Charles Lu, Anastasios N. Angelopoulos, and Stuart Pomerantz. 2022. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees.

Tomoko Nemoto and David Beglar. 2014. Likert-scale questionnaires. In *JALT 2013 conference proceedings*, pages 1–8.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling.

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

Yaniv Romano, Evan Patterson, and Emmanuel Candes. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust llm judgments? reliability of llm-as-a-judge.

Matteo Sesia and Emmanuel J. Candès. 2019. A comparison of some conformal quantile regression methods. *Stat*, 9.

Matteo Sesia and Yaniv Romano. 2021. Conformal prediction using conditional histograms.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Yishen Song, Qianta Zhu, Huaibo Wang, and Qinhua Zheng. 2024. Automated essay scoring and revising based on open-source large language models. *IEEE Transactions on Learning Technologies*, 17:1880–1890.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. Api is enough: Conformal prediction for large language models without logit-access.

Ting Fang Tan, Kabilan Elangovan, Liyuan Jin, Yao Jie, Li Yong, Joshua Lim, Stanley Poh, Wei Yan Ng, Daniel Lim, Yuhe Ke, Nan Liu, and Daniel Shu Wei Ting. 2024. Fine-tuning large language model (llm) artificial intelligence chatbots in ophthalmology and llm-based evaluation using gpt-4.

Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. 2022. Mapie: an open-source library for distribution-free uncertainty quantification.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback.

Gerrit J. J. van den Burg, Gen Suzuki, Wei Liu, and Murat Sensoy. 2025. Aligning black-box language models with human judgments.

Harit Vishwakarma, Alan Mishler, Thomas Cook, Niccolo Dalmasso, Natraj Raman, and Sumitra Ganesh. 2025. Prune 'n predict: Optimizing LLM decision-making with conformal prediction. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Nico Wagner, Michael Desmond, Rahul Nair, Zahra Ashktorab, Elizabeth M. Daly, Qian Pan, Martín Santillán Cooper, James M. Johnson, and Werner Geyer. 2024. Black-box uncertainty quantification method for llm-as-a-judge.

Victor Wang, Michael J. Q. Zhang, and Eunsol Choi. 2025. Improving llm-as-a-judge inference with the judgment distribution.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Hengtao Shen, and Xiaofeng Zhu. 2024. Conu: Conformal uncertainty in large language models with correctness coverage guarantees.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. 2025. Rocketeval: Efficient automated llm evaluation via grading checklist.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. 2025. An empirical analysis of uncertainty in large language model evaluations.

Ran Xie, Rina Foygel Barber, and Emmanuel J. Candès. 2024. Boosted conformal prediction intervals.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024a. Sayself: Teaching llms to express confidence with self-reflective rationales.

Yunpeng Xu, Wenge Guo, and Zhi Wei. 2024b. Conformal risk control for ordinal classification.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words?

Javier Yong, Haokai Ma, Yunshan Ma, Anis Yusof, Zhenkai Liang, and Ee-Chien Chang. 2025. Attackseqbench: Benchmarking large language models' understanding of sequential patterns in cyber attacks.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the confidence of large language models by eliciting fidelity.

Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and Chen Ma. 2025. Crowd comparative reasoning: Unlocking comprehensive evaluations for llm-as-a-judge.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Han Zhou, Xingchen Wan, Yinhong Liu, Nigel Collier, Ivan Vulić, and Anna Korhonen. 2024. Fairer preferences elicit improved human-aligned large language model judgments.

A Appendix

A.1 Prompt Used in Our Analysis

In our analysis, we adopted the LLM-as-a-judge frameworks G-Eval (Liu et al., 2023) across all tasks and SocREval (He et al., 2024) specifically for reasoning tasks, making only minimal prompt adjustments to suit each evaluation. Below we provide three representative prompt examples: the relevance evaluations on SummEval, and the ROSCOE evaluations under both G-Eval and SocREval.

Prompt on Relevance of SummEval

You'll be handed a summary of a news article.

Your challenge is to rate how well the summary captures the essence of the article.

Make sure to thoroughly read and understand these instructions before diving in. Keep this guide handy as you work through the task, so you can refer back to it if needed.

Evaluation Criteria:
Relevance (1–5): Does the summary hit the mark by including the most important content from the original article? It should focus on the key details without wandering into irrelevant or repetitive information. If the summary strays or over-explains, it should be rated lower.

How to Evaluate:

1. Read both the source article and the summary attentively.
2. Compare the two, identifying the critical points of the article.
3. Judge how well the summary captures these important points and avoids unnecessary details.
4. Give the summary a relevance score between 1 and 5.

Source Article:
`{{Document}}`

Summary:
`{{Summary}}`

Evaluation Form (ENTER A SCORE BETWEEN 1–5):

Relevance:

Prompt on ROSCOE by G-Eval

You will receive a generated response based on the question.

Your mission is to assess whether the generated response answers the question in a well-justified manner.

Please pay close attention to the instructions and keep this guide handy while completing your review. Feel free to refer back to it as needed.

Evaluation Criterion:

Quality (1–5): 1=incomprehensible and wrong, 5=clear and correct

Steps for Evaluation:

1. Carefully read the Situation (Premise) and the Claim (Hypothesis) to understand their key elements.
2. Evaluate the response to determine if it logically connects the Situation to the Claim and sufficiently explains whether the Claim is supported (or not) based on the details provided.
3. Score the response's quality on a scale of 1 to 5, with 1 being the lowest and 5 the highest, based on clarity, correctness, and the justification provided.

Situation (Premise):

`{{Premise}}`

Claim (Hypothesis):

`{{Hypothesis}}`

Question:

`{{Question}}`

Generated response:

`{{Response}}`

Evaluation Form (Only provide the overall quality score from 1–5 without any text):

Quality:

It is worth noting that, through oversight, we applied the SummEval prompt template directly to the DialSumm task. Even though DialSumm is a dataset for dialogue summarization rather than news summarization, the resulting intervals on DialSumm still exhibited great performance.

Prompt on ROSCOE by SocREval

Instruction:

Does the generated response answer the question in a well-justified manner? Please generate your own response for the question first, then conduct a qualitative analysis on the generated response by taking into account your own response, and finally give me an overall quality score in [1, 2, 3, 4, 5] (1=incomprehensible and wrong, 5=clear and correct). Note that you need to consider both the explanation and the answer in the generated response.

Example:

- **Situation (Premise):** Two women are embracing while holding to go packages.
- **Claim (Hypothesis):** Two women are holding packages.
- **Question:** Is the Claim supported by the Situation?
- **Generated response:** The two women are most likely embracing because they are either friends or family. If they were just holding packages, there would be no need for them to embrace. The answer is Yes.
- **Example representation:**
  - "own response": "Yes, the Claim is supported by the Situation. The Situation states that two women are ëmbracing while holding to go packages,ẅhich means they are holding packages. Therefore, the Claim is accurate.",
  - "qualitative analysis": "The generated response starts by making an assumption about why the two women might be embracing, suggesting they could be friends or family. While this could be true, it is not directly relevant to the question asked. The core of the question is about the Claim's accuracy given the Situation. The latter part of the generated response does correctly determine that the answer is Ẏes.Ḣowever, the earlier assumption makes the response longer and potentially more confusing.",
  - "overall quality": 4

Actual Input:

- **Situation (Premise):** `{{Premise}}`
- **Claim (Hypothesis):** `{{Hypothesis}}`
- **Question:** `{{Question}}`
- **Generated response:** `{{Response}}`

Evaluation Request:

Please generate your own response for the question first, then conduct a qualitative analysis on the generated response by taking into account your own response, and finally give me the overall quality of the given generated response for the question by taking into account both your own response and the qualitative analysis based on the instruction and the format of the example representation.

Evaluation Form (Only provide the overall quality score from 1–5 without any text):

Quality:

A.2 Proof of Boundary Adjustment Non-decreasing Coverage Guarantee

Proof. By the standard split conformal prediction procedure with the nonconformity score $s(x, y) = |\hat{y} - y|$, the prediction set

$$\mathcal{C}(x_{test}) = \{z \in \mathbb{R} : s(x_{test}, z) \leq \hat{q}_{1-\alpha}\}$$

satisfies

$$\mathbb{P}(Y_{test} \in \mathcal{C}(x_{test})) \geq 1 - \alpha.$$

In our discrete setting, every potential label is an element of a predetermined ordered set (e.g., $\{1, 2, 3, 4, 5\}$). The adjusted score $s'(x, y)$ is defined such that for each y,

$$s'(x, y) = s\Big(x, y'\Big),$$

where y' is the label nearest to y from the appropriate side.

In regions where the original interval $\mathcal{C}(x_{test}) = [l, u]$ already contains some labels, the shrinking adjustment leads to

$$s'(x_{test}, l) = s\Big(x_{test}, \lceil l \rceil\Big) \leq \hat{q}_{1-\alpha}$$

or

$$s'(x_{test}, u) = s\Big(x_{test}, \lfloor u \rfloor\Big) \leq \hat{q}_{1-\alpha}.$$

Thus, every label that was originally covered (i.e., satisfying $s(x_{test}, y) \leq \hat{q}$) remains covered, ensuring that the coverage remains unchanged.

On the other hand, suppose that an expanding adjustment is performed, we have

$$s'(x_{test}, l) \leq \hat{q}_{1-\alpha} \leq s'\Big(x_{test}, \lfloor l \rfloor\Big) \leq \hat{q}_{1-\alpha_0}$$

or

$$s'(x_{test}, u) \leq \hat{q}_{1-\alpha} \leq s'\Big(x_{test}, \lceil u \rceil\Big) \leq \hat{q}_{1-\alpha_0},$$

where $0 \leq \alpha_0 < \alpha$.

In this case, for any $z \notin \mathcal{C}(x_{test})$, it is possible that $z \in \mathcal{C}'(x_{test})$ for z could be $\lfloor l \rfloor$ or $\lceil u \rceil$. As a consequence, if the original interval barely missed covering the label, the expansion guarantees that these outcomes are now covered.

Hence, the event

$$\{Y_{test} \in \mathcal{C}(x_{test})\} \subseteq \{Y_{test} \in \mathcal{C}'(x_{test})\},$$

which implies

$$\mathbb{P}\Big(Y_{test} \in \mathcal{C}'(x_{test})\Big) \geq \mathbb{P}\Big(Y_{test} \in \mathcal{C}(x_{test})\Big) \geq 1-\alpha.$$

Moreover,

$$\begin{aligned}&\mathbb{P}\Big(Y_{\text{test}} \in \mathcal{C}'(x_{\text{test}})\Big) - \mathbb{P}\Big(Y_{\text{test}} \in \mathcal{C}(x_{\text{test}})\Big)\\ &= \mathbb{P}\Big(q_{1-\alpha} \leq s'\Big(x_{test}, \lfloor l \rfloor \text{ or } \lceil u \rceil\Big) \leq q_{1-\alpha_0}\Big)\\ &= (1-\alpha_0) - (1-\alpha) = \alpha - \alpha_0 > 0.\end{aligned}$$

Thus

$$\mathbb{P}\Big(Y_{test} \in \mathcal{C}'(x_{test})\Big) > 1 - \alpha.$$

□

A.3 Hypothesis Testing of Heteroskedasticity

In order to assess the validity of regression-based conformal prediction and to guide our choice of conformal prediction (CP) methods, we perform two classical tests for heteroskedasticity: the Breusch–Pagan (BP) test and the White test. This phenomenon indicates a non-constant residual variance, which causes deviation in coverage rates and inefficiency in interval widths. This data property also motivates the development of modern CP algorithms such as CQR (Romano et al., 2019), LCP (Guan, 2022), and R2CCP (Guha et al., 2024).

Breusch–Pagan Test. The BP test regresses the squared OLS residuals $\hat{e}_i^2$ on the original covariates X. Under the null hypothesis of homoskedasticity,

$$H_0 : \mathrm{Var}(\varepsilon_i) = \sigma^2 \quad \text{vs.} \quad H_1 : \mathrm{Var}(\varepsilon_i) = \sigma^2 h(X_i),$$

the test statistic

$$\mathrm{LM}_{BP} \;=\; n\, R^2_{\hat{e}^2 \sim X} \;\dot{\sim}\; \chi^2_k,$$

where n is the sample size and $k = \dim(X)$. A small p-value indicates rejection of homoskedasticity.

White Test. The White test extends BP by including not only X but also their squares and pairwise interactions $Z = \{X,\, X^2,\, X_i X_j\}$ in the auxiliary regression of $\hat{e}_i^2$. The statistic

$$\mathrm{LM}_{White} \;=\; n\, R^2_{\hat{e}^2 \sim Z} \;\dot{\sim}\; \chi^2_m,$$

with $m = \dim(Z)$. Unlike BP, White's method does not require specifying the form of $h(\cdot)$.

Test Results. Table 4 reports both BP and White p-values across our datasets and evaluators.

- **SummEval / DialSumm (G-Eval)**: All four metrics and all models exhibit highly significant heteroskedasticity ($p < 10^{-12}$).

- **ROSCOE by G-Eval**: CosmosQA remains homoskedastic, whereas DROP, e-SNLI and GSM8k show $p < 0.05$.
- **ROSCOE by SocREval**: Heteroskedasticity is confined to DROP (for DSR1-Qwen-32B and GPT-4o-mini) and to CosmosQA/e-SNLI (for Qwen2.5-72B).

A.4 Summary of CP Methods

In our analysis, we employ a total of seven regression-based conformal prediction (CP) methods to generate continuous confidence intervals (e.g., [3.2,4.1]), as well as two ordinal classification-based CP methods to produce ordered discrete intervals (e.g., [3,4]). In this subsection, we provide a detailed discussion of these CP approaches, including the motivation behind our choice to focus on regression and ordinal formulations rather than commonly used risk-control-based methods. We further elaborate on how each method computes nonconformity scores and constructs predictive intervals accordingly.

A.4.1 Why not use classification methods?

As mentioned, early work has primarily applied conformal prediction to classification-style tasks, which produces non-ordered prediction set, e.g. {A, C} in multiple choice question answering. Admittedly, the rating scale {1,2,3,4,5} can be cast as a multiple-choice classification problem. However, it is unclear how to interpret a predicted set such as {1,5}: what does it mean for both the lowest and highest scores to be both plausible, and nothing in between? As Wang et al. (2025) have shown, judgment distributions from LLMs can be irregular or even bimodal, making such fragmented prediction sets not only difficult to interpret, but also problematic for downstream decision-making.

In contrast, regression-based and ordinal conformal predictors generate ordered prediction intervals, offering a coherent and interpretable depiction of score variability. These intervals communicate not just inclusion, but range—what is the highest plausible score, and what is the lowest? In high-stakes applications such as medical diagnosis, this becomes crucial. For example, if an LLM evaluator assigns a rating of 3 (e.g., "moderate condition"), a disjoint set like {1,5} offers confusing insight. On the other hand, a calibrated interval such as [3,5] conveys that the case might be severe, thus signaling the need for a more cautious and proactive treatment plan.

A.4.2 Continuous CP methods

The following gives a brief description of each CP method used in our experiments, including its nonconformity score, interval construction procedure and how we employ.

Conformalized Quantile Regression (CQR) (Romano et al., 2019)

- *Nonconformity score:*

$$s_i = \max\{\hat{q}_{\alpha/2}(x_i)-y_i,\ y_i-\hat{q}_{1-\alpha/2}(x_i)\},$$

  where $\hat{q}_\tau$ is the τ–quantile regression estimator.
- *Interval construction:* Compute s_i on calibration set and let $Q_{1-\alpha}$ be the $(1-\alpha)$-quantile of $\{s_i\}$. For a test input x, form

$$\big[\hat{q}_{\alpha/2}(x) - Q_{1-\alpha},\ \hat{q}_{1-\alpha/2}(x) + Q_{1-\alpha}\big].$$

- *Deployment:* We implement Conformalized Quantile Regression (CQR) using the `MapieQuantileRegressor` from the `mapie` package (Taquet et al., 2022), with a `GradientBoostingRegressor` (configured for quantile loss) as the base estimator for quantile regression.

Asymmetric CQR (Sesia and Candès, 2019)

- *Nonconformity scores:*

$$s_i^\ell = \hat{q}_\alpha(x_i) - y_i, \quad s_i^u = y_i - \hat{q}_{1-\alpha}(x_i).$$

- *Interval construction:* Let Q_ℓ and Q_u be the $(1-\alpha)$-quantiles of $\{s_i^\ell\}$ and $\{s_i^u\}$, respectively. Then

$$\big[\hat{q}_\alpha(x) - Q_\ell,\ \hat{q}_{1-\alpha}(x) + Q_u\big].$$

- *Deployment:* Same with CQR but the asymmetric variant.

Conditional Histogram Regression (CHR) (Sesia and Romano, 2021)

- *Distribution estimation:* Partition the target range into bins and estimate $\Pr(Y \in \text{bin} \mid X = x)$ via a black-box model.
- *Nested set series:* Based on conditional probability, construct a series of nested set $\{C_t\}_{t=0}^T$, where T is the length of the series and C_t expands as t increases.

SummEval by G-Eval																	
Evaluator	**Test**	Consistency				Coherence				Fluency				Relevance			
		LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value
GPT-4o-mini	BP	372.121	3.02e-78	96.615	4.52e-89	147.034	5.71e-30	32.261	2.07e-31	144.954	1.58e-29	31.759	6.35e-31	102.860	1.32e-20	21.903	2.85e-21
	White	446.359	4.68e-82	30.547	2.61e-97	204.285	1.60e-32	11.556	4.96e-35	187.021	4.08e-29	10.450	3.71e-31	132.282	1.45e-18	7.116	1.85e-19
DSR1-Qwen-32B	BP	332.234	1.17e-69	83.545	4.44e-78	64.602	1.36e-12	13.414	7.81e-13	209.266	2.95e-43	47.970	2.46e-46	78.494	1.73e-15	16.447	7.41e-16
	White	406.728	8.21e-74	26.910	4.32e-86	142.666	1.58e-20	7.729	1.33e-21	242.606	3.52e-40	14.111	6.53e-44	92.448	2.76e-11	4.841	1.19e-11
Qwen2.5-72B	BP	351.775	7.26e-74	89.844	2.03e-83	82.248	2.84e-16	17.276	1.11e-16	227.917	2.99e-47	52.956	5.94e-51	83.830	1.32e-16	17.627	4.96e-17
	White	407.695	5.17e-74	26.996	2.33e-86	142.134	1.99e-20	7.697	1.71e-21	245.423	9.55e-41	14.304	1.40e-44	100.688	9.49e-13	5.302	3.34e-13

DialSumm by G-Eval																	
Evaluator	**Test**	Consistency				Coherence				Fluency				Relevance			
		LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value
GPT-4o-mini	BP	70.220	9.22e-14	14.723	4.30e-14	199.050	4.54e-41	46.209	2.85e-44	250.633	4.02e-52	60.796	2.06e-57	87.825	1.92e-17	18.664	5.49e-18
	White	96.250	5.87e-12	5.091	2.01e-12	238.533	2.32e-39	14.160	1.56e-43	271.824	4.40e-46	16.613	9.80e-52	170.231	7.82e-26	9.548	9.86e-28
DSR1-Qwen-32B	BP	100.158	4.90e-20	21.483	9.27e-21	126.174	1.54e-25	27.616	9.87e-27	169.680	8.54e-35	38.451	4.63e-37	177.728	1.64e-36	40.540	5.12e-39
	White	169.039	1.33e-25	9.468	1.83e-27	196.532	5.45e-31	11.260	1.21e-33	225.735	8.54e-37	13.255	1.83e-40	250.758	8.03e-42	15.045	1.65e-46
Qwen2.5-72B	BP	88.782	1.21e-17	18.877	3.37e-18	209.551	2.57e-43	49.076	6.76e-47	199.737	3.23e-41	46.395	1.92e-44	125.827	1.83e-25	27.532	1.19e-26
	White	123.892	5.40e-17	6.694	7.13e-18	228.974	1.92e-37	13.482	3.09e-41	235.628	8.89e-39	13.953	7.83e-43	175.737	6.61e-27	9.897	6.01e-29

ROSCOE by G-Eval																	
Evaluator	**Test**	CosmosQA				DROP				e-SNLI				GSM8k			
		LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value
GPT-4o-mini	BP	5.839	0.3222	1.167	0.3270	11.334	0.0451	2.328	0.0440	26.074	0.0001	6.053	0.0000	7.586	0.1806	1.530	0.1822
	White	17.194	0.6404	0.841	0.6609	23.456	0.2669	1.188	0.2681	35.174	0.0192	1.974	0.0124	26.151	0.1609	1.346	0.1556
DSR1-Qwen-32B	BP	8.042	0.1539	1.626	0.1550	20.313	0.0011	4.369	0.0008	24.209	0.0002	5.537	0.0001	15.828	0.0074	3.335	0.0065
	White	17.670	0.6092	0.867	0.6290	40.833	0.0039	2.281	0.0022	58.598	0.0000	4.122	0.0000	33.872	0.0270	1.825	0.0210
Qwen2.5-72B	BP	7.883	0.1628	1.592	0.1641	22.042	0.0005	4.785	0.0004	22.554	0.0004	5.092	0.0002	27.782	0.0000	6.259	0.0000
	White	25.904	0.1690	1.333	0.1640	31.326	0.0510	1.657	0.0438	49.770	0.0002	3.196	0.0000	56.739	0.0000	3.545	0.0000

ROSCOE by SocREval																	
Evaluator	**Test**	CosmosQA				DROP				e-SNLI				GSM8k			
		LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value	LM Stat.	p-value	F Stat.	p-value
GPT-4o-mini	BP	7.256	0.20231	1.461	0.20457	4.016	0.54705	0.796	0.55399	7.637	0.17742	1.545	0.17954	3.577	0.61180	0.707	0.61918
	White	20.130	0.26762	1.199	0.26973	37.301	0.01077	2.041	0.00732	12.766	0.75172	0.722	0.77546	11.661	0.82022	0.663	0.83626
DSR1-Qwen-32B	BP	5.659	0.34085	1.130	0.34606	29.404	0.00002	6.643	0.00001	6.994	0.22105	1.409	0.22457	4.244	0.51487	0.841	0.52199
	White	13.283	0.86492	0.636	0.88162	38.105	0.00860	2.095	0.00561	18.955	0.52477	0.933	0.54694	8.983	0.98311	0.421	0.98678
Qwen2.5-72B	BP	13.470	0.01935	2.805	0.01810	8.464	0.13245	1.714	0.13293	16.545	0.00545	3.569	0.00450	2.321	0.80313	0.456	0.80886
	White	34.356	0.00755	2.227	0.00499	22.780	0.19917	1.291	0.19706	29.926	0.03818	1.813	0.02981	12.387	0.77613	0.707	0.79345

Table 4: Breusch-Pagan (BP) and White tests detect pervasive heteroscedasticity in SummEval and DialSumm: both tests yield highly significant p-values ($p < 1e-12$) across all four metrics and all evaluators. By contrast, in ROSCOE by G-Eval only DROP, e-SNLI and GSM8k exhibit significant heteroscedasticity ($p < 0.05$) while CosmosQA remains homoscedastic; in ROSCOE by SocREval heteroscedasticity is confined to DROP for DeepSeek-R1-Qwen-32B and GPT-4omini and to CosmosQA and e-SNLI for Qwen2.5-72B-Instruct.

- *Compute conformity score*

$$s_i = \min\{t \in \{0, ..., T\} : y \in C_t\}$$

  on calibration set and obtain estimated quantile $s_{1-\alpha} = \hat{t}$.
- *Interval construction:* Find the $\hat{t}$-th set $C_{\hat{t}}$ in $\{C_t(x_{test})\}_{t=0}^{T}$ for the test point.
- *Deployment:* We estimate the conditional distributions by QNet estimator with two hidden layers of 256 units each, a batch size of 32, learning rate 5×10^{-4} and 1000 epochs.

Locally Valid and Discriminative (LVD) (Lin et al., 2021)

- *Nonconformity score:* Absolute residuals $R_i = |y_{n+i} - \hat{y}_{n+i}|$ are collected as nonconformity scores from the calibration set. Here, $\hat{y}_{n+i}$ can be the prediction from the a deep neural network model or from kernel regression.
- *Interval construction:* For a test point, weights are assigned to the calibration residuals based on its similarity to points in the calibration set, quantified by a trained kernel function K_f. The $(1-\alpha)$-quantile of this weighted residual distribution is then calculated. This quantile is added to and subtracted from the base deep learning model's prediction $\hat{y}_{N+1}$ for the test point, yielding locally adaptive prediction intervals.

Locally Valid and Discriminative (LVD) (Lin et al., 2021)

- *Nonconformity scores:* For each calibration example (x_{n+i}, y_{n+i}), compute the absolute residual

$$R_i = \big|\, y_{n+i} - \hat{y}_{n+i} \big|,$$

  where $\hat{y}_{n+i}$ is the model's point prediction (e.g. from a deep network or kernel regression).

- *Interval construction:* For a test input x, assign similarity weights

$$w_i \propto K_f\big(x_{n+i},\, x\big) \quad \text{and} \quad w_\infty \propto K_f(x, x)$$

(normalized so $\sum_i w_i + w_\infty = 1$), form the weighted empirical distribution of $\{R_i\}$ with a "safe-guard" atom at ∞, take its $(1-\alpha)$–quantile Q, and output

$$\big[\hat{y}(x) - Q,\ \hat{y}(x) + Q\big].$$

- *Deployment:* We train the kernel similarity function using KernelMLKR with parameters `d=10, seed=0, n_iters=500, norm=True, lr=1e-3` that used in their demo notebook.

Boosted Conformal Prediction (Xie et al., 2024)

- *Boosting the conformity score:* BoostedCP optimizes conformity score functions from baselines like CQR (Romano et al., 2019) or LCP (Guan, 2022) via gradient boosting. This is guided by a tailored loss function, aiming for enhanced conditional coverage or reduced interval length. It operates post-model training, solely relying on model predictions.
- *Interval Estimation:* The boosted score function is used for calibration to compute empirical quantiles. These quantiles, with the boosted score, construct final intervals for testing points. This approach improves prediction interval statistical properties while maintaining valid marginal coverage.
- *Deployment:* We set `n_rounds_cv` as 500 and `learning_rate` as 0.02.

R2CCP (Regression-to-Classification Conformal Prediction) (Guha et al., 2024)

- *Two-stage approach:* Partition the continuous response range $[y_{\min}, y_{\max}]$ into K equally spaced bins with midpoints $\{\hat{y}_k\}_{k=1}^{K}$. Train a softmax-output neural network to classify.
- *Non-conformity score:* On the calibration set, compute for each pair the interpolated probability $\sigma_j = \bar{q}_\theta(y_j \mid x_j)$ by linearly interpolating q_θ between adjacent bin midpoints.
- *Interval construction:* Obtain $(1-\alpha)$-quantile $\hat{q}_{1-\alpha}$ of non-conformity scores and generate intervals by

$$C_{1-\alpha}(x_{test}) = \{\, z \in \mathbb{R} : \bar{q}_\theta(z \mid x) \geq \hat{q}_{1-\alpha} \,\}.$$

- *Deployment:* We train R2CCP model with `max_epoches = 100`. In practice, this method might yields fragmental intervals. We merge those intervals into one by taking minimum and maximum. Moreover, the range of labels in calibration determines the bin split in testing. Thus there would be error if two ranges are inconsistent, which causes that the trials of random experiments are somesometimes slightly less than 30.

A.4.3 Ordinal CP methods

Ordinal CP methods generate intervals by softmax probabilities, which derives from judgment distribution. For GPA-scale tasks, we obtain probabilities of fractional labels (e.g. 1.33, 1.67, ...) by linear interpolation.

Ordinal APS (Lu et al., 2022)

- *Nonconformity score:* Nonconformity score equals to 1 if the true label lies in the interval and 0 if not. Obtain an empirical quantile λ as the threshold of accumulated probability mass.
- *Interval construction:* Start from the label with highest probability, and then extend to both directions until the accumulated probability mass reach the quantile.

Ordinal Risk Control (Xu et al., 2024b)

- *Nonconformity score:* Similar to Ordinal APS but calculate the empirical risk by weighted average of nonconformity scores. Select a smallest quantile λ to control the empirical risk.
- *Interval Estimation:* Similar to Ordinal APS, start from point estimation and extend to both directions until the miscoverage risk is higher than λ.
- *Deployment:* We deploy the `WeightedCRPredictor` variant for better performance in our tasks.

A.5 Human-based baseline in summarization tasks

Table 5 demonstrates that R2CCP consistently matches or outperforms the human baseline across both SummEval and DialSumm.

A.6 In Context Learning G-Eval

Following ICE (Jain et al., 2023) and G-Eval (Liu et al., 2023), we designed the prompts by example selection with three sampling methods. For each test sample, we randomly select examples with difference source test as the exsample pool. There are 100 source in the SummEval dataset, so the sample size of example pool is $99 \times 16 = 1584$.

As for sampling method, ICE (Jain et al., 2023) has introduced uniform sampling and stratified sampling to in-context LLM evaluation on SummEval. We modified the later to quantile-based sampling to stratify bins by distribution quantiles. For example, assume we need K examples in a prompt, uniform sampling is to randomly select K samples in the example pool, stratified sampling is to stratify the range of scores into K bins and then randomly select 1 from each bin, and quantile based sampling is to stratify the distribution of scores into K quantile bins and select 1 from each bin.

After prompt design, we obtained evaluations from GLM 4-flash (GLM et al., 2024), which is free to use API. Then we calculate the correlation with the expert average. Here we present several results of different number of shots and different sampling methods, comparing with results of G-Eval and ICE. We found that ICL-G-Eval based on GLM 4-flash is equivalent to GPT3.5-Eval.

Adding more samples leads to varying effects on correlation across dimensions. For coherence and fluency, the impact is minimal or slightly negative. In contrast, consistency and relevance benefit, particularly under the quantile method. Among evaluation methods, quantile performs best in relevance, while stratified excels in other three dimensions.

A.7 Supplementary Results and Analysis

A.7.1 Continuous Intervals

In Table 7, we observe that some methods such as Boosted CQR and Boosted LCP consistently fall short of the 90 % coverage target on the DialSumm dataset, achieving only 86%–88%. In contrast, R2CCP maintains coverage in the 89%–91% range while yielding the narrowest intervals among methods with comparable performance, thus offering an optimal trade-off between coverage and efficiency. LVD achieves slightly higher coverage (around 90 %–92 %) but at the cost of wider intervals, making it suitable for scenarios that prioritize coverage over interval compactness. Both Asymmetric and Symmetric CQR reliably guarantee or exceed the 90 % coverage, but the cost is larger interval widths (mostly larger than 3 on ROSCOE). Across evaluators, we find that the intervals produced by the DSR1-Distill-Qwen-32B model achieve marginally higher average coverage rates. And those generated by Qwen2.5-72B-Instruct are generally shorter with lower coverage.

A.7.2 Discrete Intervals

Overall, with the aid of boundary adjustment, nearly all continuous-interval methods achieve average coverage rates of approximately 90% (Table 8). With coverage guarantee, the gap between BoostedCP (Boosted CQR and Boosted LCP) and R2CCP become narrower, all of which now offer similarly optimal trade-offs between coverage and interval width.

When comparing across evaluators, we observe heterogeneity in method performance: on GPT-4o mini, Boosted CQR and Boosted LCP typically attain the best balance, with R2CCP slightly behind; in contrast, under DSR1-Distill-Qwen-32B and Qwen2.5-72B-Instruct, the ordinal methods (OrdinalAPS, OrdinalRC), which generally underperform on GPT-4o mini, excel on DSR1. Notably, OrdinalAPS produces interval with markedly smaller intervals than those of the other methods while the coverage is around 90%.

A.7.3 With Only 0.1 Adjustment is Effective to Mitigate Miscoverage

In Section 3.3, we introduced boundary adjustment, whereby a continuous confidence interval on a Likert scale is rounded to the nearest integer endpoints, i.e. any true label falling within half an ordinal step of a boundary is adjusted to that boundary. In practice, we observed that purely continuous intervals sometimes underperform the nominal 90% coverage target (Table 1 and 7), owing to the heteroskedastic, and correlated nature of LLM-generated judgments and to calibration set sizes that are insufficient. Full boundary adjustment reliably remedies this miscoverage by converting continuous intervals into ordinal discrete intervals, but it may introduce bias or fail to satisfy users' preference for continuous outputs. To strike

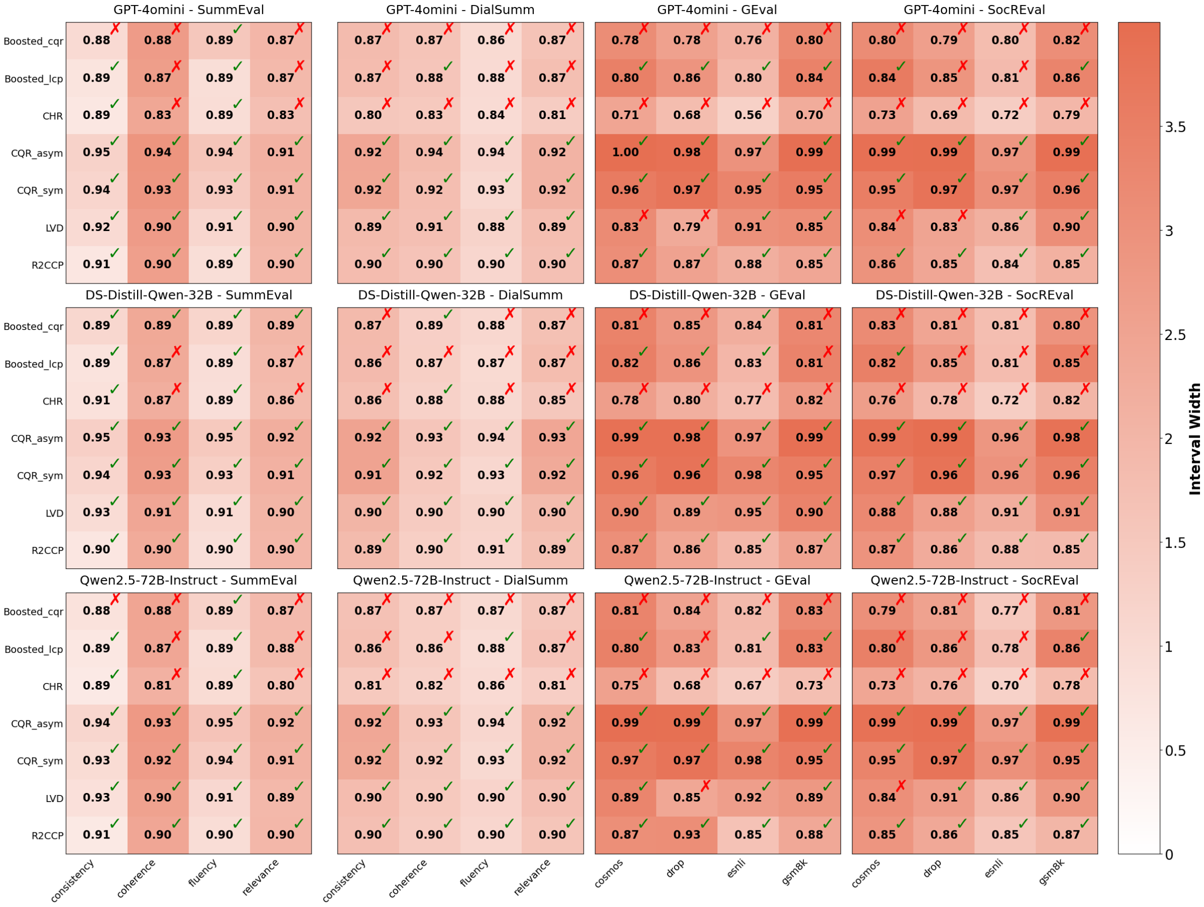

Figure 5: This is the summary and comparison on all experiments for continuous intervals. Each cell displays the mean coverage of its corresponding CP method over 30 trials on the given dataset. The cell's shading encodes the average interval width, with lighter hues denoting narrower intervals. A ✓ or ✗ in the cell's upper-right corner denotes whether the coverage criterion is met: specifically, if the mean coverage plus one standard deviation exceeds 90%, the cell is marked with ✓; otherwise, it is marked with ✗.

Dataset	Evaluator	Method	Metrics			
			Consistency	**Coherence**	**Fluency**	**Relevance**
SummEval	Human-based	Baseline	0.667 (91.4%)	2.000 (95.6%)	1.333 (96.3%)	2.000 (92.8%)
	GPT-4o-mini	R2CCP	0.621 (90.1%)	2.652 (89.9%)	1.135 (93.4%)	2.076 (91.5%)
	DSR1-Qwen-32B	R2CCP	0.598 (89.3%)	2.168 (85.8%)	0.850 (90.1%)	2.142 (93.3%)
	Qwen-2.5-72B	R2CCP	0.491 (88.9%)	2.429 (88.0%)	0.812 (88.0%)	1.969 (91.4%)
DialSumm	Human-based	Baseline	2.667 (95.9%)	2.000 (96.9%)	2.000 (95.1%)	2.667 (95.6%)
	GPT-4o-mini	R2CCP	1.799 (91.99%)	1.701 (91.00%)	1.215 (89.71%)	1.580 (85.2%)
	DSR1-Qwen-32B	R2CCP	1.912 (88.7%)	1.283 (89.3%)	0.812 (88.0%)	1.805 (89.9%)
	Qwen-2.5-72B	R2CCP	1.591 (87.0%)	1.494 (90.3%)	1.136 (90.3%)	1.653 (91.9%)

Table 5: Comparison of human-based baseline and R2CCP (seed = 42) on SummEval and DialSumm

Metric	Coherence		Consistency		Fluency		Relevance	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ
G-EVAL-3.5	0.440	0.335	0.386	0.318	0.424	0.347	0.401	0.320
G-EVAL-4	**0.582**	**0.457**	**0.507**	**0.425**	**0.547**	**0.433**	**0.514**	**0.418**
ICE (Uniform Sampling)	0.476	**0.388**	**0.486**	**0.466**	0.366	0.328	0.467	0.384
ICE (Stratified Sampling)	**0.497**	0.387	0.298	0.263	**0.397**	**0.348**	**0.485**	**0.396**
ICL-G-Eval 0-shot	0.411	0.354	**0.441**	**0.417**	**0.404**	**0.384**	0.428	**0.375**
ICL-G-Eval 1-shot uniform	0.411	0.339	0.405	0.374	0.302	0.282	0.406	0.341
ICL-G-Eval 2-shot quantile	0.445	0.365	0.422	0.387	0.345	0.321	0.439	0.367
ICL-G-Eval 3-shot stratified	**0.447**	**0.367**	0.424	0.385	0.356	0.335	0.430	0.361
ICL-G-Eval 4-shot stratified	0.409	0.334	0.421	0.382	0.362	0.340	0.412	0.340
ICL-G-Eval 5-shot stratified	0.393	0.323	0.419	0.381	0.356	0.334	0.414	0.339
ICL-G-Eval 5-shot quantile	0.417	0.343	0.404	0.364	0.302	0.280	**0.448**	**0.375**

Table 6: Summary-level Spearman and Kendall-Tau correlations of different metrics on the SummEval benchmark

a balance, we propose a partial boundary adjustment with threshold λ (e.g. $\lambda = 0.1$), meaning that only those interval endpoints within λ of an integer are rounded. For instance, [3.2,4.9] becomes [3,5] under full adjustment, but under $\lambda = 0.1$ it becomes [3.2,5], which increases coverage if the true label is 5.

As our theorem certifies, this outward adjustment effectively shifts the quantile levels to include more potential labels within the interval. Empirically, larger λ yields greater coverage gains, while the average interval width does not increase too much and can even shrink. This is because our adjustment simultaneously cut redundant fractional parts that fail to cover the true label (e.g. [1.05,2.1] becomes [1,2], removing the excessive [2,2.1] segment). If the frequencies and sums of shrinking and expanding adjustments across intervals are approximately balanced, the average interval width remains unchanged. A sufficient condition for this result is that, within each integer bin, boundaries' fractional parts (e.g. 0.3 of 4.3, 0.7 of 1.7) are symmetrically distributed. But due to unknown distributions of model output, formally verifying this theorem remains challenging.

A.7.4 Midpoints

Table 3, 11, 12 and 13 show that midpoints are less-biased score evaluations than baselines from LLM judgments.

A.7.5 Reprompt and Regrade

Table 14 shows that there is trivial difference if we try to reprompt LLM judges to regrade. However, Figure 6, 7, 8, 9, 10, 11 show the examples of responses in reprompting.

Evaluator	Method	DialSumm (G-Eval)				ROSCOE (G-Eval)			
		Consistency	Coherence	Fluency	Relevance	CosmosQA	DROP	e-SNLI	GSM8K
GPT-4o mini	Boosted CQR	1.85 / 86.81%	1.61 / 87.26%	1.03 / 86.33%	1.65 / 87.06%	3.12 / 77.99%	2.58 / 78.32%	2.13 / 75.79%	3.20 / 80.03%
	Boosted LCP	1.83 / 87.45%	1.59 / 88.30%	1.00 / 87.53%	1.76 / 87.20%	3.45 / 79.66%	2.94 / 86.41%	1.94 / 80.26%	3.42 / 83.53%
	CHR	1.54 / 80.01%	1.48 / 83.03%	0.99 / 84.01%	1.40 / 80.53%	2.47 / 70.78%	1.82 / 68.44%	1.28 / 55.66%	2.27 / 70.27%
	Asym CQR	2.43 / 92.30%	1.87 / 94.00%	1.18 / 94.40%	2.09 / 92.38%	3.95 / 99.56%	3.89 / 98.19%	**2.98 / 96.67%**	3.94 / 99.27%
	Sym CQR	2.41 / 91.99%	1.77 / 92.41%	**1.08 / 93.38%**	2.06 / 91.57%	**3.60 / 96.43%**	**3.77 / 96.54%**	3.35 / 95.31%	**3.58 / 94.83%**
	LVD	1.90 / 89.20%	1.75 / 90.67%	1.20 / 88.40%	1.79 / 89.23%	3.18 / 83.44%	2.33 / 79.11%	3.00 / 91.18%	3.10 / 84.50%
	R2CCP	**1.84 / 90.13%**	**1.63 / 90.15%**	1.14 / 89.64%	**1.72 / 90.11%**	3.09 / 86.77%	2.54 / 86.70%	2.20 / 88.01%	2.43 / 84.67%
DSR1-Distill-Qwen-32B	Boosted CQR	1.89 / 87.48%	1.31 / 88.61%	1.11 / 88.07%	1.71 / 87.39%	3.40 / 80.92%	2.84 / 85.02%	2.23 / 83.68%	3.27 / 80.53%
	Boosted LCP	1.88 / 86.05%	1.32 / 86.77%	1.02 / 87.28%	1.82 / 87.24%	3.49 / 81.73%	2.94 / 86.06%	1.99 / 83.33%	3.38 / 81.13%
	CHR	1.76 / 86.09%	1.26 / 87.80%	1.06 / 87.79%	1.53 / 85.22%	2.64 / 78.10%	2.19 / 80.13%	1.80 / 77.24%	2.77 / 81.90%
	Asym CQR	2.52 / 91.92%	1.58 / 93.20%	1.22 / 94.04%	2.42 / 92.51%	3.89 / 98.95%	3.88 / 97.78%	2.95 / 97.15%	3.90 / 99.27%
	Sym CQR	2.50 / 91.30%	1.51 / 91.85%	**1.11 / 92.83%**	2.33 / 91.69%	**3.62 / 96.29%**	**3.82 / 96.22%**	3.33 / 97.85%	3.54 / 95.27%
	LVD	**2.03 / 90.19%**	**1.41 / 90.29%**	1.22 / 90.41%	**1.87 / 90.01%**	3.31 / 89.52%	2.81 / 88.98%	**2.86 / 94.82%**	**3.39 / 90.07%**
	R2CCP	1.86 / 89.22%	1.31 / 89.92%	1.19 / 90.57%	1.70 / 89.39%	3.05 / 86.84%	2.44 / 85.87%	1.96 / 85.43%	2.51 / 86.77%
Qwen2.5-72B-Instruct	Boosted CQR	1.69 / 86.57%	1.35 / 87.06%	1.05 / 87.38%	1.52 / 87.08%	3.35 / 81.02%	2.55 / 83.52%	1.90 / 81.84%	3.18 / 82.70%
	Boosted LCP	1.75 / 86.08%	1.35 / 86.29%	0.96 / 87.77%	1.63 / 86.88%	3.45 / 80.41%	2.79 / 83.05%	1.85 / 80.79%	3.42 / 83.47%
	CHR	1.48 / 81.50%	1.25 / 81.88%	1.00 / 85.97%	1.33 / 81.15%	2.59 / 74.97%	1.76 / 68.38%	1.39 / 66.84%	1.76 / 72.57%
	Asym CQR	2.40 / 91.92%	1.61 / 92.88%	1.14 / 93.87%	2.04 / 92.38%	3.93 / 99.12%	3.92 / 99.17%	2.96 / 97.41%	3.90 / 99.00%
	Sym CQR	2.37 / 91.51%	**1.52 / 91.57%**	**1.06 / 93.11%**	**2.02 / 91.55%**	**3.62 / 96.67%**	3.78 / 96.86%	3.37 / 98.29%	**3.58 / 95.30%**
	LVD	**1.84 / 90.43%**	1.47 / 89.91%	1.20 / 90.26%	1.74 / 89.99%	3.34 / 89.01%	2.41 / 84.57%	**2.65 / 92.41%**	2.83 / 88.50%
	R2CCP	1.74 / 89.97%	1.41 / 89.67%	1.14 / 89.70%	1.61 / 89.80%	3.07 / 87.11%	**3.10 / 93.40%**	1.68 / 84.80%	2.38 / 87.53%

Table 7: Comparison of interval width and coverage across conformal methods on DialSumm and ROSCOE tasks with G-Eval. Gray marks coverage <85%, underline marks coverage between 85%–90%, and **bold** highlights the smallest interval width among methods achieving $\geq$90% coverage for each evaluator–dimension. Asymmetric CQR yields the highest coverage but with wider intervals; R2CCP and LVD can offer narrower intervals while still meeting the coverage target, making them preferable when efficiency matters.

Evaluator	Method	DialSumm (G-Eval)				ROSCOE (G-Eval)			
		Consistency	Coherence	Fluency	Relevance	CosmosQA	DROP	e-SNLI	GSM8K
GPT-4o mini	Boosted CQR	1.85 / 93.33%	1.60 / 93.95%	1.00 / 93.32%	**1.66 / 92.66%**	3.16 / 93.40%	2.60 / 90.51%	2.16 / 89.39%	**3.22 / 90.50%**
	Boosted LCP	**1.83 / 92.94%**	1.60 / 93.01%	**0.96 / 93.50%**	1.76 / 91.85%	3.39 / 94.46%	2.97 / 91.71%	**1.96 / 92.32%**	3.32 / 92.87%
	CHR	1.54 / 86.65%	**1.47 / 90.28%**	0.96 / 89.86%	1.40 / 87.49%	2.48 / 80.31%	1.82 / 78.06%	1.30 / 72.41%	2.25 / 78.90%
	Asym CQR	2.43 / 94.34%	1.86 / 95.87%	1.18 / 95.97%	2.08 / 94.68%	3.95 / 99.69%	3.89 / 98.54%	2.99 / 97.54%	3.94 / 99.27%
	Sym CQR	2.40 / 94.09%	1.76 / 94.50%	1.07 / 95.01%	2.05 / 94.33%	3.60 / 96.60%	3.77 / 96.86%	3.42 / 97.50%	3.57 / 95.10%
	LVD	1.90 / 93.81%	1.75 / 94.43%	1.21 / 93.81%	1.80 / 93.77%	3.20 / 91.70%	2.33 / 86.63%	3.01 / 96.89%	3.11 / 89.53%
	R2CCP	1.84 / 93.32%	1.63 / 93.38%	1.15 / 93.28%	1.72 / 93.65%	**3.06 / 90.31%**	**2.52 / 90.48%**	2.16 / 92.35%	2.42 / 86.77%
	OrdinalAPS	2.25 / 56.79%	2.03 / 82.21%	1.87 / 48.44%	2.07 / 66.38%	1.79 / 70.61%	1.44 / 78.57%	1.75 / 70.13%	1.36 / 75.03%
	OrdinalRC	2.33 / 91.49%	3.17 / 39.46%	2.01 / 64.66%	2.21 / 83.29%	1.94 / 73.16%	1.52 / 80.73%	1.84 / 72.32%	1.44 / 75.47%
DSR1-Distill-Qwen-32B	Boosted CQR	1.89 / 92.91%	1.31 / 94.36%	1.08 / 94.16%	1.71 / 92.95%	3.44 / 94.69%	2.88 / 94.41%	2.27 / 95.44%	3.29 / 93.60%
	Boosted LCP	1.87 / 91.83%	1.32 / 93.31%	0.98 / 93.27%	1.82 / 91.52%	3.49 / 94.76%	3.03 / 91.27%	2.01 / 92.59%	3.31 / 91.70%
	CHR	**1.76 / 90.78%**	1.25 / 93.09%	1.03 / 92.48%	1.53 / 90.65%	2.66 / 85.95%	2.21 / 87.94%	1.83 / 90.57%	2.75 / 89.17%
	Asym CQR	2.51 / 93.90%	1.58 / 95.41%	1.22 / 95.61%	2.43 / 94.35%	3.89 / 98.95%	3.88 / 97.87%	2.94 / 97.28%	3.90 / 99.37%
	Sym CQR	2.49 / 93.67%	1.51 / 94.80%	1.11 / 94.58%	2.31 / 93.54%	3.61 / 96.43%	3.83 / 96.95%	3.32 / 97.98%	3.54 / 95.80%
	LVD	2.04 / 93.87%	1.41 / 95.07%	1.23 / 94.74%	1.87 / 93.86%	3.34 / 94.69%	2.82 / 93.40%	2.87 / 98.55%	3.42 / 95.47%
	R2CCP	1.85 / 92.87%	1.31 / 93.84%	1.19 / 93.79%	1.70 / 93.23%	3.04 / 91.29%	2.40 / 89.84%	**1.90 / 90.79%**	2.49 / 88.87%
	OrdinalAPS	0.68 / 89.98%	**0.90 / 90.52%**	**0.93 / 90.51%**	**0.68 / 90.28%**	**2.90 / 90.99%**	**2.27 / 91.24%**	3.20 / 91.93%	2.98 / 91.93%
	OrdinalRC	2.05 / 90.04%	3.17 / 90.30%	3.42 / 89.93%	2.17 / 89.80%	2.79 / 89.59%	2.22 / 90.73%	3.15 / 90.48%	**2.86 / 90.83%**
Qwen2.5-72B-Instruct	Boosted CQR	1.70 / 92.85%	**1.35 / 93.90%**	1.05 / 93.82%	1.52 / 92.95%	3.40 / 94.56%	2.57 / 93.30%	1.92 / 94.47%	3.22 / 92.03%
	Boosted LCP	1.76 / 92.50%	**1.35 / 93.38%**	**0.90 / 92.39%**	1.62 / 92.52%	3.45 / 95.24%	2.85 / 90.95%	**1.91 / 92.02%**	3.38 / 92.73%
	CHR	1.48 / 87.99%	1.25 / 89.46%	0.97 / 91.59%	1.34 / 88.06%	2.62 / 83.74%	1.77 / 82.92%	1.43 / 85.31%	1.78 / 82.10%
	Asym CQR	2.41 / 95.10%	1.61 / 95.21%	1.14 / 95.08%	2.04 / 95.00%	3.93 / 99.42%	3.92 / 99.24%	2.95 / 97.50%	3.89 / 99.07%
	Sym CQR	2.37 / 94.62%	1.51 / 94.42%	1.06 / 94.40%	2.05 / 94.91%	3.63 / 97.14%	3.78 / 97.14%	3.38 / 98.42%	3.58 / 95.73%
	LVD	1.84 / 94.48%	1.48 / 94.77%	1.20 / 95.10%	1.73 / 94.04%	3.36 / 94.90%	**2.41 / 92.44%**	2.65 / 98.25%	2.83 / 92.47%
	R2CCP	1.73 / 93.55%	1.41 / 93.72%	1.15 / 93.83%	1.60 / 93.17%	3.05 / 90.71%	3.08 / 95.65%	1.59 / 89.67%	2.39 / 89.60%
	OrdinalAPS	**0.61 / 90.15%**	0.87 / 78.65%	0.91 / 85.01%	**0.69 / 90.18%**	2.78 / 89.42%	**2.02 / 90.95%**	2.79 / 93.25%	**2.46 / 90.30%**
	OrdinalRC	2.57 / 89.85%	3.01 / 68.81%	3.09 / 78.40%	2.81 / 89.81%	**2.85 / 90.95%**	1.93 / 89.62%	2.71 / 91.40%	2.60 / 91.63%

Table 8: Comparison of narrow discrete intervals and coverage across methods on DialSumm and ROSCOE with G-Eval. Gray marks coverage $<$ 85%, underline marks coverage between 85%–90%, and **bold** highlights the smallest interval width among methods achieving $\geq$ 90% coverage for each evaluator–dimension. Comparing with Table 7, we could find that all coverage rates improve due to boundary adjustment, while the interval widths remain comparable. For ordinal CP methods, since they directly produce discrete intervals, boundary adjustment has no influence to their intervals for the bounaries are already on the potential labels$\{1.00, 1.33, ..., 4.67, 5.00\}$.

Model	Judge	Dataset	0.5	0.1	0
GPT-4o-mini	G-Eval	Cosmos	3.0612 ± 0.5594 / 90.31% ± 7.07%	3.0847 ± 0.5371 / 87.31% ± 8.17%	3.0864 ± 0.5335 / 86.77% ± 7.98%
		DROP	2.5230 ± 0.4804 / 90.48% ± 5.53%	2.5410 ± 0.4402 / 87.30% ± 6.03%	2.5431 ± 0.4363 / 86.70% ± 5.87%
		e-SNLI	2.1562 ± 0.4732 / 92.35% ± 6.88%	2.1932 ± 0.4282 / 88.30% ± 7.15%	2.1953 ± 0.4264 / 88.01% ± 7.17%
		GSM8K	2.4205 ± 0.7782 / 86.77% ± 7.63%	2.4283 ± 0.7639 / 85.10% ± 7.50%	2.4298 ± 0.7626 / 84.67% ± 8.00%
	SocREval	Cosmos	2.9294 ± 0.4597 / 89.46% ± 7.01%	2.9586 ± 0.4396 / 86.73% ± 7.80%	2.9618 ± 0.4350 / 85.85% ± 7.79%
		DROP	2.4125 ± 0.8208 / 89.21% ± 9.21%	2.4271 ± 0.7631 / 85.40% ± 10.04%	2.4300 ± 0.7600 / 84.73% ± 9.97%
		e-SNLI	1.7076 ± 0.5804 / 90.11% ± 8.41%	1.7467 ± 0.4878 / 84.99% ± 8.23%	1.7480 ± 0.4842 / 84.02% ± 8.62%
		GSM8K	2.0943 ± 1.1782 / 86.93% ± 8.15%	2.1452 ± 1.0893 / 85.70% ± 7.95%	2.1452 ± 1.0866 / 85.07% ± 7.87%
DSR1-Distill-Qwen-32B	G-Eval	Cosmos	3.0357 ± 0.4873 / 91.29% ± 5.71%	3.0474 ± 0.4629 / 87.35% ± 6.04%	3.0489 ± 0.4601 / 86.84% ± 5.87%
		DROP	2.4000 ± 0.6406 / 89.84% ± 7.80%	2.4360 ± 0.5746 / 86.48% ± 7.94%	2.4385 ± 0.5714 / 85.87% ± 8.05%
		e-SNLI	1.8952 ± 0.4948 / 90.79% ± 7.39%	1.9532 ± 0.4343 / 86.06% ± 6.95%	1.9585 ± 0.4315 / 85.43% ± 7.02%
		GSM8K	2.4865 ± 0.8454 / 88.87% ± 9.31%	2.5067 ± 0.8065 / 87.07% ± 9.04%	2.5078 ± 0.8053 / 86.77% ± 8.89%
	SocREval	Cosmos	2.9094 ± 0.6323 / 90.58% ± 7.92%	2.9365 ± 0.5718 / 87.76% ± 8.45%	2.9378 ± 0.5689 / 86.97% ± 8.37%
		DROP	2.2457 ± 0.6021 / 89.97% ± 8.61%	2.2906 ± 0.5369 / 86.92% ± 9.06%	2.2931 ± 0.5342 / 86.35% ± 9.08%
		e-SNLI	1.7965 ± 0.5139 / 92.35% ± 7.77%	1.8413 ± 0.4481 / 88.45% ± 8.00%	1.8450 ± 0.4443 / 87.87% ± 7.92%
		GSM8K	1.8238 ± 1.2189 / 86.93% ± 7.45%	1.8767 ± 1.1507 / 85.67% ± 7.30%	1.8796 ± 1.1480 / 85.33% ± 7.02%
Qwen2.5-72B-Instruct	G-Eval	Cosmos	3.0529 ± 0.5262 / 90.71% ± 6.80%	3.0624 ± 0.5089 / 87.82% ± 8.11%	3.0652 ± 0.5059 / 87.11% ± 8.10%
		DROP	3.0765 ± 0.9169 / 95.65% ± 5.33%	3.0954 ± 0.8907 / 93.68% ± 7.50%	3.0964 ± 0.8894 / 93.40% ± 7.74%
		e-SNLI	1.5885 ± 0.4282 / 89.67% ± 6.42%	1.6737 ± 0.3734 / 85.28% ± 6.56%	1.6792 ± 0.3694 / 84.80% ± 6.89%
		GSM8K	2.3922 ± 0.6387 / 89.60% ± 4.34%	2.3778 ± 0.6328 / 87.93% ± 5.71%	2.3782 ± 0.6323 / 87.53% ± 5.95%
	SocREval	Cosmos	2.8786 ± 0.5572 / 89.29% ± 7.43%	2.8999 ± 0.5210 / 86.16% ± 8.76%	2.8996 ± 0.5176 / 85.34% ± 8.46%
		DROP	2.3446 ± 0.6459 / 90.00% ± 7.90%	2.3832 ± 0.5883 / 86.92% ± 8.14%	2.3852 ± 0.5854 / 86.25% ± 8.22%
		e-SNLI	1.5461 ± 0.6282 / 90.20% ± 8.42%	1.5854 ± 0.5467 / 85.43% ± 8.44%	1.5897 ± 0.5397 / 84.50% ± 8.75%
		GSM8K	1.9602 ± 1.0970 / 88.57% ± 7.01%	2.0026 ± 1.0422 / 87.07% ± 7.14%	2.0015 ± 1.0391 / 86.73% ± 7.09%

Table 9: R2CCP interval width and coverage under boundary adjustments 0.5, 0.1, and 0 for three models, two judge frameworks, and four reasoning datasets (width±std / coverage%± std) based on 30 random trials: all coverages improve.

Model	Dataset	Dimension	0.167 (Full Adjustment)	0.1	0
GPT-4o-mini	SummEval	Consistency	0.6753 ± 0.2026 / 92.15% ± 2.25%	0.6800 ± 0.1951 / 91.68% ± 2.33%	0.6858 ± 0.1859 / 90.88% ± 2.49%
		Coherence	2.6186 ± 0.1522 / 92.81% ± 2.37%	2.6201 ± 0.1497 / 91.54% ± 2.69%	2.6243 ± 0.1466 / 89.63% ± 3.12%
		Fluency	0.9116 ± 0.1673 / 90.99% ± 2.06%	0.9166 ± 0.1657 / 90.49% ± 2.29%	0.9213 ± 0.1641 / 89.36% ± 2.71%
		Relevance	1.9688 ± 0.1288 / 93.38% ± 1.96%	1.9693 ± 0.1244 / 91.90% ± 2.19%	1.9705 ± 0.1215 / 89.70% ± 2.50%
	DialSumm	Consistency	1.8443 ± 0.1299 / 93.32% ± 1.85%	1.8425 ± 0.1298 / 92.03% ± 1.96%	1.8436 ± 0.1287 / 90.13% ± 2.39%
		Coherence	1.6264 ± 0.1363 / 93.38% ± 2.68%	1.6274 ± 0.1354 / 92.10% ± 3.08%	1.6256 ± 0.1337 / 90.15% ± 3.38%
		Fluency	1.1504 ± 0.1187 / 93.28% ± 2.03%	1.1484 ± 0.1237 / 91.87% ± 2.56%	1.1357 ± 0.1226 / 89.64% ± 2.92%
		Relevance	1.7161 ± 0.1398 / 93.65% ± 2.06%	1.7178 ± 0.1391 / 92.15% ± 2.32%	1.7209 ± 0.1395 / 90.11% ± 2.75%
DeepSeek-R1-Distill-Qwen-32B	SummEval	Consistency	0.6804 ± 0.1521 / 91.57% ± 2.17%	0.6876 ± 0.1437 / 91.02% ± 2.11%	0.6941 ± 0.1343 / 90.44% ± 2.09%
		Coherence	2.2972 ± 0.1161 / 93.22% ± 1.65%	2.2994 ± 0.1169 / 91.91% ± 1.88%	2.3042 ± 0.1172 / 90.12% ± 2.13%
		Fluency	0.8886 ± 0.1605 / 91.80% ± 1.82%	0.8907 ± 0.1561 / 91.09% ± 2.00%	0.8926 ± 0.1512 / 90.09% ± 2.08%
		Relevance	1.9935 ± 0.1557 / 92.96% ± 2.09%	1.9951 ± 0.1514 / 91.72% ± 2.35%	1.9984 ± 0.1482 / 89.84% ± 2.90%
	DialSumm	Consistency	1.8534 ± 0.1426 / 92.87% ± 2.01%	1.8574 ± 0.1397 / 91.43% ± 2.10%	1.8601 ± 0.1371 / 89.22% ± 2.62%
		Coherence	1.3113 ± 0.1082 / 93.84% ± 1.85%	1.3126 ± 0.1077 / 92.28% ± 2.14%	1.3138 ± 0.1068 / 89.92% ± 2.76%
		Fluency	1.1903 ± 0.1368 / 93.79% ± 1.78%	1.1915 ± 0.1349 / 92.58% ± 2.03%	1.1859 ± 0.1348 / 90.57% ± 2.35%
		Relevance	1.6952 ± 0.1660 / 93.23% ± 1.89%	1.6982 ± 0.1639 / 91.70% ± 2.16%	1.7043 ± 0.1601 / 89.39% ± 2.57%
Qwen2.5-72B-Instruct	SummEval	Consistency	0.5876 ± 0.1520 / 91.83% ± 1.92%	0.5973 ± 0.1447 / 91.47% ± 1.88%	0.6122 ± 0.1341 / 90.73% ± 2.02%
		Coherence	2.4308 ± 0.1457 / 92.78% ± 2.07%	2.4331 ± 0.1444 / 91.53% ± 2.19%	2.4367 ± 0.1426 / 89.54% ± 2.48%
		Fluency	0.9494 ± 0.1180 / 92.12% ± 1.46%	0.9500 ± 0.1216 / 91.38% ± 1.76%	0.9527 ± 0.1218 / 90.17% ± 1.92%
		Relevance	1.9765 ± 0.1257 / 93.72% ± 1.78%	1.9776 ± 0.1253 / 92.50% ± 2.06%	1.9789 ± 0.1237 / 90.45% ± 2.49%
	DialSumm	Consistency	1.7319 ± 0.1106 / 93.55% ± 1.59%	1.7350 ± 0.1083 / 92.16% ± 1.91%	1.7368 ± 0.1050 / 89.97% ± 2.18%
		Coherence	1.4060 ± 0.1115 / 93.72% ± 1.97%	1.4079 ± 0.1086 / 92.07% ± 2.36%	1.4094 ± 0.1076 / 89.67% ± 2.81%
		Fluency	1.1518 ± 0.1265 / 93.83% ± 2.29%	1.1475 ± 0.1345 / 92.37% ± 2.74%	1.1376 ± 0.1398 / 89.70% ± 3.42%
		Relevance	1.5966 ± 0.1742 / 93.17% ± 2.02%	1.6015 ± 0.1714 / 91.82% ± 2.43%	1.6071 ± 0.1682 / 89.80% ± 2.85%

Table 10: R2CCP interval width and coverage under boundary adjustments of 0.5, 0.1, and 0 for three models on SummEval and DialSumm across four dimensions (width±std / coverage%±std): all coverages improve.

Evaluator	Method	Coherence				Consistency				Fluency				Relevance			
		MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ
GPT-4omini	Raw Score	3.787	1.711	0.205	0.172	1.000	0.772	0.656	0.547	2.111	1.171	0.400	0.344	1.278	0.874	0.668	0.564
	Weighted Sum	3.701	1.699	0.218	0.162	0.825	0.704	0.702	0.546	1.688	1.066	0.434	0.338	1.175	0.855	0.703	0.549
	OrdinalAPS	**0.834**	**0.733**	**0.287**	**0.236**	2.284	1.358	0.685	**0.572**	2.358	1.327	0.414	**0.345**	1.802	1.137	0.639	0.524
	Con_R2CCP	**0.344**	**0.454**	**0.396**	**0.300**	**0.391**	**0.489**	0.688	0.532	**0.173**	**0.309**	0.433	0.340	**0.338**	**0.445**	**0.716**	0.563
	Dis_R2CCP	**0.348**	**0.453**	**0.385**	**0.313**	**0.395**	**0.489**	0.684	**0.553**	**0.178**	**0.309**	0.418	**0.364**	**0.342**	**0.446**	**0.714**	**0.584**
DSR1-Qwen-32B	Raw Score	2.908	1.412	0.396	0.329	1.422	0.952	0.589	0.497	2.454	1.383	0.414	0.356	1.214	0.829	0.555	0.461
	Weighted Sum	2.149	1.241	0.456	0.343	0.652	0.614	0.642	0.491	2.115	1.287	0.452	0.347	0.674	0.625	0.621	0.476
	OrdinalAPS	**0.124**	**0.276**	**0.476**	**0.411**	**0.078**	**0.215**	0.641	**0.552**	**0.127**	**0.297**	0.367	0.325	**0.092**	**0.237**	0.555	0.474
	Con_R2CCP	**0.211**	**0.348**	**0.627**	**0.488**	**0.451**	**0.509**	**0.668**	**0.512**	**0.185**	**0.315**	**0.460**	0.356	**0.348**	**0.450**	**0.721**	**0.563**
	Dis_R2CCP	**0.215**	**0.347**	**0.615**	**0.511**	**0.455**	**0.508**	**0.665**	**0.534**	**0.188**	**0.314**	**0.455**	**0.389**	**0.352**	**0.450**	**0.716**	**0.581**
Qwen2.5-72B	Raw Score	3.934	1.775	0.321	0.267	1.344	0.897	0.704	0.599	2.796	1.420	0.478	0.406	1.812	1.070	0.609	0.521
	Weighted Sum	3.693	1.746	0.358	0.266	1.076	0.819	0.737	0.577	2.575	1.335	0.499	0.386	1.552	1.014	0.660	0.516
	OrdinalAPS	**0.192**	**0.379**	**0.436**	**0.385**	**0.080**	**0.209**	0.709	**0.617**	**0.133**	**0.293**	0.398	0.352	**0.105**	**0.244**	0.610	**0.528**
	Con_R2CCP	**0.241**	**0.381**	**0.583**	**0.450**	**0.370**	**0.467**	0.737	0.577	**0.169**	**0.306**	0.489	0.380	**0.311**	**0.424**	**0.727**	**0.578**
	Dis_R2CCP	**0.245**	**0.380**	**0.574**	**0.471**	**0.373**	**0.467**	0.734	0.594	**0.174**	**0.306**	0.485	**0.419**	**0.315**	**0.424**	**0.722**	**0.594**

Table 11: Midpoints experiment on DialSumm: The midpoints substantially reduce MSE and MAE while boosting Spearman's ρ and Kendall's τ across all evaluators and dimensions, outperforming both Raw Score and Weighted Sum. **Bold** indicates better performance than baselines, underlined denotes comparable performance, and gray indicates worse performance.

Evaluator	Method	CosmosQA				DROP				e-SNLI				GSM8k			
		MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ
GPT-4o-mini	Raw Score	1.780	1.044	0.483	0.406	1.843	0.951	0.490	0.411	2.719	1.210	0.340	0.288	2.216	0.909	0.586	0.516
	Weighted Sum	1.704	1.065	0.490	0.371	1.651	0.894	0.516	0.391	2.610	1.221	0.357	0.273	2.169	0.936	0.577	0.458
	OrdinalAPS	1.719	1.087	0.466	0.385	**1.374**	**0.869**	0.523	**0.432**	**2.141**	**1.142**	**0.358**	**0.290**	**1.970**	1.001	0.585	0.495
	Con_R2CCP	2.035	1.223	0.366	0.282	**1.509**	1.034	0.458	0.353	**1.045**	**0.865**	0.239	0.189	2.307	1.282	0.493	0.396
	Dis_R2CCP	2.044	1.220	0.348	0.293	**1.526**	1.024	0.469	0.402	**1.061**	**0.854**	0.231	0.206	2.317	1.277	0.501	0.434
DSR1-Qwen-32B	Raw Score	2.353	1.166	0.396	0.335	2.156	0.977	0.478	0.419	3.090	1.466	0.225	0.199	2.300	0.906	0.596	0.538
	Weighted Sum	1.805	1.157	0.462	0.348	1.281	0.913	0.551	0.422	2.144	1.286	0.279	0.214	1.907	1.045	0.602	0.476
	OrdinalAPS	1.836	1.163	0.433	**0.366**	1.305	0.916	**0.558**	**0.464**	**1.874**	**1.185**	0.268	**0.233**	2.247	1.387	0.554	0.497
	Con_R2CCP	1.931	1.172	0.440	0.344	1.485	1.003	0.491	0.380	**0.904**	**0.802**	**0.423**	**0.334**	2.232	1.283	0.540	0.432
	Dis_R2CCP	1.936	**0.999**	0.407	0.345	1.518	0.999	0.478	0.406	**0.916**	**0.792**	**0.405**	**0.355**	2.256	1.281	0.539	0.472
Qwen2.5-72B	Raw Score	1.964	1.179	0.420	0.364	1.797	0.928	0.498	0.421	1.920	1.173	0.359	0.304	1.911	0.820	0.653	0.589
	Weighted Sum	1.840	1.166	0.484	0.367	1.381	0.867	0.569	0.437	1.615	1.101	0.388	0.292	1.767	0.857	0.662	0.529
	OrdinalAPS	**1.775**	**1.158**	0.460	**0.385**	**1.183**	**0.852**	**0.582**	**0.485**	**1.391**	**1.001**	**0.484**	**0.414**	1.816	1.168	0.639	0.561
	Con_R2CCP	1.992	1.207	0.429	0.329	1.789	1.124	**0.584**	**0.453**	**0.796**	**0.727**	**0.471**	**0.372**	2.021	1.174	0.566	0.460
	dis_R2CCP	2.004	1.200	0.390	0.327	1.801	1.121	**0.573**	**0.487**	**0.816**	**0.716**	**0.455**	**0.400**	2.044	1.173	0.578	0.511

Table 12: Midpoints experiment on ROSCOE evaluated by G-Eval. **Bold** indicates better performance than baselines, underlined denotes comparable performance, and gray indicates worse performance.

Evaluator	Method	CosmosQA				DROP				e-SNLI				GSM8k			
		MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ	MSE	MAE	ρ	τ
GPT-4o-mini	Raw Score	1.780	1.044	0.483	0.406	2.969	1.284	0.202	0.168	1.096	0.841	0.551	0.496	4.103	1.613	0.173	0.148
	Weighted Sum	1.704	1.065	0.490	0.371	1.408	0.905	0.563	0.429	1.054	0.849	0.574	0.460	1.612	0.771	0.649	0.523
	OrdinalAPS	1.719	**1.034**	0.466	**0.412**	1.447	0.958	0.555	**0.449**	1.108	0.880	0.567	0.486	1.672	0.794	**0.654**	**0.571**
	Con_R2CCP	1.904	1.170	0.430	0.330	1.560	1.017	0.495	0.386	**0.725**	**0.716**	0.509	0.408	2.061	1.154	0.569	0.470
	Dis_R2CCP	1.917	1.165	0.415	0.348	1.578	1.013	0.493	0.421	**0.753**	**0.711**	0.505	0.453	2.095	1.144	0.589	0.527
DSR1-Qwen-32B	Raw Score	2.130	1.128	0.500	0.432	1.443	0.803	0.630	0.564	0.693	0.629	0.581	0.531	1.445	0.628	0.707	0.640
	Weighted Sum	2.016	1.107	0.525	0.398	1.446	0.825	0.639	0.503	0.668	0.632	0.622	0.496	1.425	0.645	0.664	0.522
	OrdinalAPS	**1.862**	1.128	0.484	0.399	**1.305**	0.900	0.573	0.508	**0.663**	0.653	**0.631**	**0.555**	**1.395**	0.830	0.696	0.628
	Con_R2CCP	**1.853**	1.151	0.468	0.362	**1.264**	0.914	0.602	0.476	0.717	0.708	0.615	0.490	1.849	1.048	0.595	0.477
	Dis_R2CCP	**1.875**	1.146	0.595	0.515	**1.290**	0.907	0.595	**0.515**	0.734	0.695	0.580	0.517	1.891	1.045	0.637	0.577
Qwen2.5-72B	Raw Score	1.737	0.975	0.533	0.444	1.313	0.730	0.610	0.536	0.590	0.488	0.651	0.591	1.387	0.653	0.730	0.663
	Weighted Sum	1.688	0.986	0.527	0.407	1.290	0.735	0.603	0.475	0.558	0.499	0.665	0.540	1.388	0.659	0.681	0.557
	OrdinalAPS	**1.622**	0.991	**0.542**	**0.450**	**1.281**	0.744	**0.610**	**0.526**	**0.543**	**0.543**	**0.650**	**0.575**	1.390	0.667	0.727	0.657
	Con_R2CCP	1.897	1.162	0.446	0.348	1.378	0.968	0.556	0.456	0.609	0.656	0.597	0.482	1.823	1.063	0.648	0.556
	dis_R2CCP	1.910	1.156	0.442	0.375	1.403	0.965	0.541	0.487	0.632	0.652	0.595	0.533	1.849	1.057	0.653	0.596

Table 13: Midpoints experiment on ROSCOE evaluated by SocREval. **Bold** indicates better performance than baselines, underlined denotes comparable performance, and gray indicates worse performance.

Reprompt on ROSCOE by DeepSeek-R1-Distill-Qwen-32B with SocREval

Let me show you our evaluation record. Based on all these information, make decision and give me final score.

Initial Prompt:
`{{Prompt in 1st round}}`

Initial Response:
`{{Response in 1st round}}`

Reprompt and Regrade:

Thank you for your initial evaluation!

To help you arrive at a final score that more closely aligns with human expert judgment, we have constructed a 90% confidence interval for this task using conformal prediction based on your past scoring records. This interval is provided to help you gauge the uncertainty in your recent assessment, which we hope will enhance your evaluation.

Interval Information: The confidence interval we have provided is `{{Interval}}`. Please keep in mind that there is approximately a 90% probability that the expert's score lies within this interval, and a 10% probability that it lies outside.

— **Your Objective**: Acting as a human expert, use the interval information along with the recent evaluation task to decide whether and how to adjust the initial score.

— Below are some decision-making suggestions for your reference, but we also encourage you to apply your own independent thinking to align as closely as possible with human expert judgment.

Decision-Making Suggestions:

1. **Key Dimensions to Consider**:
    - **Original Score Confidence**: Your confidence level in the score you just assigned;
    - **Interval Position**: Whether the original score falls inside or outside the interval;
    - **Interval Width**: Whether the interval is narrow (e.g., ≤ 2.0) or wide (e.g., ≥ 2.0);
    - **Potential Labels**: What specific label options lie within the interval (e.g., {3.00, 4.00} for an example interval [3,4]).
2. **Advice on Decision, Reasoning, and Explanation**:
    - **High Confidence & Score Within a Narrow Interval**: If the interval is narrow and your score is validated by the interval boundaries, you may confidently retain your original score, provided you believe your evaluation and explanation are seamless. You may also make minor adjustments within the interval where you think the score is most plausible.
    - **High Confidence & Score Outside a Narrow Interval**: Although there is a small probability (<10%) that you are correct, we encourage you to question your initial judgment, reconsider the evaluation, and consider adjusting the score to the most probable point within the interval, or retain the original score with a brief justification.
    - **Low Confidence & Any Interval**: Use the interval to guide a careful re-examination of the task. For example, consider why an expert rating might take a certain value within the interval and whether that reasoning is sound. After reflection, if you find a value most reasonable, you may choose that score.
    - **Challenging the Interval**: You have the right to firmly believe that the true score cannot possibly fall within the provided interval. However, since we guarantee that the interval covers the expert score 90% of the time, your challenge likely indicates an error in expert judgment. If your explanation convinces us, this would be a valuable discovery. Generally, though, we prefer to treat the expert judgment as ground truth.

Please use the suggestions above to produce a **new final score** through a step-by-step chain of thought:
1. Your confidence level in the original score (high/medium/low) and the reason;
2. How the interval width, potential labels and the position of the original score influence your judgment;
3. Your adjustment action (retain/minor adjustment/re-examination/other) and the rationale;
4. The final score you assign.

Finally, please first state your final evaluated score (1–5), followed by your explanation:

Final Score:

Figure 6: In our reprompting, the dialogue in 1st round are fed into the 2nd-round re-evaluation, and the judge is supplied with explicit guidance on how to leverage the interval for decision-making. For example, verbalize its confidence, assessing the initial score's position relative to the interval, and even challenge the interval's reliability. This prompt induces DeepSeek-R1-Distill-Qwen-32B to emulate a human-like reasoning process during inference, thereby allowing us to analyze how interval evaluations influence final judgments.

Dataset	Width / Coverage	Method	MSE	MAE	ρ	τ
CosmosQA	2.60 / 89.80%	Initial Raw	2.204082	1.163265	0.480293	0.419364
		Reprompt Raw	2.193877	1.153061	0.476310	0.417798
		Initial Weighted	2.052884	1.133203	0.508314	0.390947
		Reprompt Weighted	2.111918	1.167847	0.499106	0.377264
DROP	1.67 / 89.52%	Initial Raw	1.371429	0.800000	0.603949	0.551028
		Reprompt Raw	1.380952	0.809524	0.603821	0.550921
		Initial Weighted	1.333399	0.800079	0.612075	0.485937
		Reprompt Weighted	1.345206	0.814889	0.634605	0.503989
e-SNLI	1.26 / 89.47%	Initial Raw	0.684211	0.631579	0.561363	0.517585
		Reprompt Raw	0.657895	0.631579	0.595460	0.548320
		Initial Weighted	0.610842	0.623154	0.639802	0.512257
		Reprompt Weighted	0.605095	0.638822	0.646223	0.517462
GSM8K	1.14 / 92.00%	Initial Raw	0.860000	0.420000	0.816251	0.747567
		Reprompt Raw	0.850000	0.410000	0.819313	0.755605
		Initial Weighted	0.840141	0.437920	0.763763	0.599531
		Reprompt Weighted	0.833816	0.463369	0.738347	0.590121

Table 14: Reprompting the LLM with confidence intervals reinforces its original judgments since initial scores already lie within those intervals, the model makes trivial adjustments, revealing that it might be hard for interval alone to correct inherent bias since there is no significant difference in each metric for comparison.

Reprompt Response of LLM Judge: Ground truth is 1 in this example (DROP of ROSCOE).

Final Score: 2

Explanation:

1. **Confidence Level**: High confidence in the original score of 2. The generated response incorrectly stated the number of losses, which was a clear error.
2. **Interval Consideration**: The confidence interval [1.00, 3.00] suggests that the score is likely within this range. The original score of 2 falls within this interval, aligning with the expert judgment.
3. **Adjustment Action**: Retain the original score of 2. The reasoning remains sound as the generated response had a factual error, leading to a lower quality score.
4. **Rationale**: The generated response's error in interpreting the team's record significantly impacts its quality, justifying the score of 2.

Figure 7: Example 1 of an LLM judge's reprompted scoring and rationale. In this example, its initial raw score is 2 and initial weighted score is 1.65. After reprompting, the judge more firmly believe that 2 is right since its weighted score become 2.04, which representing a shift in judgment distributions. However, the ground truth of this is 1.

Reprompt Response of LLM Judge: Ground truth is 2 in this example (GSM8K of ROSCOE).

Final Score: 2

Explanation: After reviewing the generated response and considering the confidence interval [1.00, 2.00], I remain confident that the original score of 2 is appropriate. The generated response incorrectly calculated the water intake rate and the time taken, leading to an overestimation of the water taken on. The correct approach, as detailed in my own response, results in 16 liters, whereas the generated response arrived at 160 liters. The interval suggests that the score is likely within 1 to 2, and since the response is incorrect but shows some understanding, a score of 2 is justified.

Figure 8: Example 2 of an LLM judge's reprompted scoring and rationale. In this example, its initial raw score is 2 and initial weighted score is 1.81. After reprompting, the judge more firmly believe that 2 is right since its weighted score become 2.02, which representing a shift in judgment distributions. In this example, the ground truth is 2.

Reprompt Response of LLM Judge: Ground truth is 3 in this example (CosmosQA of ROSCOE).

Final Score: 4

Explanation: 1. Confidence in Original Score: High. The initial evaluation correctly identified that Sam's understanding supports the Claim, though it could have elaborated more on the context.
2. Interval Consideration: The interval [1.00, 5.00] is wide, allowing for a range of scores. The original score of 4 falls comfortably within this interval, suggesting it is plausible.
3. Adjustment Action: Retain. The reasoning aligns well with the provided information, and the score accurately reflects the quality of the generated response.
4. Rationale: The generated response correctly answers the question but could have provided more depth. However, it sufficiently supports the Claim, warranting a score of 4.

Figure 9: Example 3 of an LLM judge's reprompted scoring and rationale. In this example, its initial raw score is 4 and initial weighted score is 3.82. After reprompting, the judge more firmly believe that 4 is right since its weighted score become 3.97, which representing a shift in judgment distributions. In this example, the ground truth is 3.

Reprompt Response of LLM Judge: Ground truth is 5 in this example (Consistency of SummEval).

Final Score: 4

Explanation:

1. **Confidence in Original Score**: High. The initial evaluation correctly assessed the summary's accuracy without factual errors.
2. **Omitted Details**: The summary fails to mention Vanessa's profession and the police appeal, which are key elements of the source text.
3. **Interval Consideration**: The confidence interval [4.67, 5.00] indicates that a higher score is possible under the uncertainty model.
4. **Adjustment Action**: Retain. Despite the interval's suggestion, the omission of critical details justifies keeping the score at 4.

Figure 10: Example of an LLM judge's reprompted scoring and rationale. This response shows that though we guarantee the help of confidence interval, the LLM judge still insist its choice thus fails to reduce the bias.

Reprompt Response of LLM Judge: Ground truth is 5 in this example (Consistency of SummEval).

Final Score: 4

Explanation:

1. **Confidence in Original Score**: High. The summary accurately reflects the main points without introducing inaccuracies, though it omits some details.
2. **Interval Consideration**: The interval [4.67, 5.00] suggests a higher score is expected. However, the summary's omission of Pereira's contract and other clubs' interest justifies a slightly lower score.
3. **Adjustment Action**: Retain the original score of 4. The summary's focus on Orlando City's interest makes the omissions acceptable, maintaining consistency without misleading information.
4. **Final Score**: 4. The summary is truthful and consistent, though less comprehensive, warranting a score within the lower end of the interval.

Figure 11: Example of an LLM judge's reprompted scoring and rationale. This response shows that though we guarantee the help of confidence interval, the LLM judge still insist its choice thus fails to reduce the bias.