PARAPHRASE-ROBUST CONFORMAL PREDICTION FOR RELIABLE LLM UNCERTAINTY QUANTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Uncertainty quantification (UQ) provides interpretable measures of predictive confidence and supports reliable decision-making with large language models (LLMs). However, existing UQ methods are often neither statistically rigorous nor robust to paraphrase variations. To address these limitations, we propose a new framework for paraphrase-robust UQ, which builds on conformal prediction to ensure valid coverage and introduces a paraphrase-aware nonconformity score to enhance robustness. The score is derived by generating semantic paraphrases of each query, training an ancillary model that both approximates and robustifies the predictive distribution, and aggregating variability across these paraphrases. On five general multiple-choice Question Answering (MCQA) datasets and two medical MCQA datasets with `Qwen2.5-7B`, our method achieves nominal coverage with compact prediction sets and demonstrates improved robustness to paraphrase shifts in an adversarial setting. The results also generalize to `Llama-3.1-8B` and `Phi-3-small`, underscoring the reliability of the framework across model families. Code is available at `https://anonymous.4open.science/r/paraphrase_uq-FDD8`.

1 INTRODUCTION

Large language models (LLMs) have been rapidly deployed in high-stakes domains such as education and medicine (Bouchard & Chauhan, 2025; López et al., 2025). Despite their impressive performance, LLMs often exhibit overconfidence: the probabilities in their outputs do not reliably reflect the true uncertainty of their predictions (Shorinwa et al., 2025). This miscalibration poses serious risks in safety-critical applications, where decision-makers need to know how uncertain the model prediction is. However, common token-level heuristics (e.g., entropy, margins, logit ranks) provide ad hoc uncertainty estimates without statistical guarantees (Shorinwa et al., 2025; Nado et al., 2022; Ulmer et al., 2022; Band et al., 2024; Huang et al., 2024b), which limits their reliability in practice.

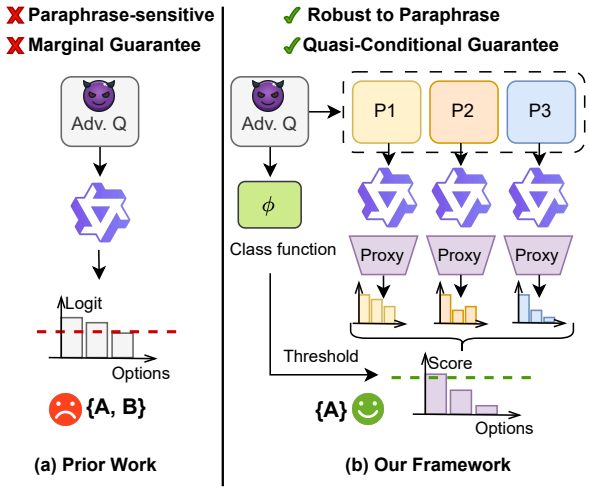

Figure 1: Comparison with prior work.

To ensure statistical validity, conformal prediction (CP) provides a principled wrapper: given any nonconformity score, CP constructs prediction sets with finite-sample coverage under mild exchangeability assumptions (Vovk et al., 2005; Shafer & Vovk, 2008). Due to this task-agnostic guarantee, CP has been applied to many LLM tasks such as multiple-choice question answering (QA), factuality evaluation, and generation alignment (Quach et al., 2024a; Gui et al., 2024; Ye et al., 2024; Wang et al., 2024b; Su et al., 2024). A key challenge, however, lies in choosing a

reliable nonconformity score. Ideally, a nonconformity score should be stable under paraphrasing: since natural language admits many equivalent expressions, such rewordings should not cause large fluctuations in the prediction set. For example, the questions "What is the capital of France?" and "Which city serves as France's capital?" should yield comparable uncertainty sets. In extreme cases, adversarial paraphrases can deliberately perturb a model's predictions while preserving semantics (see the left panel in Figure 1). Although CP still guarantees valid coverage in theory, different nonconformity scores behave very differently in finite samples, often producing unstable or inflated sets. For instance, as shown in Table 1, the prediction set size for two popular CP scores (Sadinle et al., 2019; Romano et al., 2020) nearly doubles under adversarial paraphrasing compared to clean inputs. This exposes a fundamental gap in existing CP score design, which has largely overlooked robustness to paraphrasing.

In this work, we ask whether it is possible to design a nonconformity score that remains robust under adversarial paraphrasing. We provide an affirmative answer by introducing *paraphrase-aware* nonconformity scores that explicitly enforce semantic invariance. As illustrated in the right panel of Figure 1, our pipeline generates paraphrases for each query, embeds them into a shared semantic space, and trains a lightweight proxy model to produce calibrated predictive probabilities. The resulting scores aggregate variability across paraphrases and are inherently robust to rewordings. When applied to both split CP and the finer quasi-conditional CP (QCCP) (Gibbs et al., 2025), these scores consistently preserve target coverage while substantially reducing prediction set sizes.

We summarize our contributions as follows:

- **New evaluation setting.** We introduce adversarial paraphrasing as a new setting for LLM uncertainty evaluation, and show that standard conformal methods fail under semantically equivalent rewordings.
- **Paraphrase-aware scores.** We propose *paraphrase-aware nonconformity scores* that can be applied within both split CP and QCCP, maintaining formal coverage guarantees while yielding smaller sets under adversarial paraphrasing.
- **Large-scale validation.** We conduct large-scale experiments on five general QA and two medical QA benchmarks, showing that our method consistently achieves nominal coverage with up to 2–4$\times$ smaller sets than existing baselines, together with detailed ablations and analyses.

2 Related Work

Heuristic and Calibration-Based Uncertainty for LLMs. LLM uncertainty is often estimated from token-level signals such as entropy, logits, or ranks, followed by calibration. Early work analyzes calibration of deep models and NLP tasks (Nado et al., 2022; Ulmer et al., 2022; Si et al., 2022), and extends to prompt- or generation-level schemes for long-form outputs (Band et al., 2024; Huang et al., 2024b). Recent methods replace raw probabilities with representation-based surrogates, introducing semantic entropy probes (Kossen et al., 2024), relevance-aware confidence for free-form generation (Duan et al., 2024), perturbation-based measures (Gao et al., 2024), multi-agent diversity signals (Feng et al., 2024), and semantic-density metrics (Qiu & Miikkulainen, 2024). Complementary directions include abstention mechanisms (Madhusudhan et al., 2025), multicalibration for confidence scores (Detommaso et al., 2024), post-hoc calibration from generated text (Ulmer et al., 2024), and parameter-efficient Bayesian or ensemble-style methods for fine-tuned LLMs (Balabanov & Linander, 2025; Wang et al., 2024a). While these methods improve empirical calibration, they offer neither distribution-free coverage nor robustness to paraphrasing. In contrast, our approach introduces paraphrase-aware scores that seamlessly integrates conformal prediction to provide coverage guarantees.

Conformal Prediction for LLMs. Conformal prediction (CP) provides distribution-free prediction sets with finite-sample coverage under exchangeability (Vovk et al., 2005; Shafer & Vovk, 2008). Recent work has adapted CP to language modeling in several ways, including conformal language modeling (Quach et al., 2024a;b), API-only inference without logit access (Su et al., 2024), and schemes tailored to multiple-choice questions (Kumar et al., 2023; Yang & Liu, 2025). CP has also been applied to align and certify outputs (Gui et al., 2024), as well as to benchmark LLMs with uncertainty metrics beyond accuracy (Ye et al., 2024). Extensions to generation tasks include `ConU`, which applies split CP to sets of sampled responses (Wang et al., 2024b), and `SConU`, which analyzes exchangeability violations to approximate conditional guarantees (Wang et al., 2025b).

Further refinements combine CP with re-asking strategies to improve accuracy and compactness (Vishwakarma et al., 2025), while selective answering with risk control has been explored through conformal abstention (Tayebati et al., 2025; Wang et al., 2025a). Our approach instead enforces *semantic invariance* via paraphrase-robust scores and leverages *quasi-conditional calibration* on semantic embeddings, yielding stronger coverage guarantees than the marginal coverage provided by standard split CP.

Toward Conditional Guarantees and Paraphrase Robustness. Exact conditional coverage is unattainable in finite samples without distributional assumptions (Vovk, 2012; Foygel Barber et al., 2021), which has motivated relaxations such as *quasi-conditional* guarantees via augmented quantile regression (Gibbs et al., 2025). For LLMs, recent extensions of CP exploit feature-conditional structure (Cherian et al., 2024). In parallel, another line of work emphasizes *semantic* rather than purely lexical uncertainty, introducing embedding-based metrics and perturbation procedures (Gao et al., 2024; Kossen et al., 2024; Huang et al., 2024a). Related approaches probe prompt sensitivity and meaning-preserving perturbations (Qiu & Miikkulainen, 2024; Cox et al., 2025), though they do not provide conformal guarantees. Our method integrates these two directions: (i) quantifying predictive stability across paraphrases using sentence embeddings and a proxy classifier, and (ii) calibrating these scores with QCCP (Gibbs et al., 2025), thereby producing prediction sets that are both semantically robust and statistically valid.

3 Paraphrase-Robust Quasi-Conditional Conformal Prediction

Problem Setup. LLMs often produce predictions that are brittle to paraphrasing and hard to calibrate. Our goal is to construct *prediction sets* that not only guarantee statistical coverage but also remain stable under meaning-preserving rephrasings. Formally, we consider supervised prediction with input $x \in \mathcal{X}$ (e.g., a natural language question) and a finite label space $\mathcal{Y}$ (e.g., multiple choice answers). Each example (X_i, Y_i) provides a ground-truth label $Y_i \in \mathcal{Y}$. Our goal is to construct a prediction set $\widehat{C}(x) \subseteq \mathcal{Y}$ that maintains guaranteed coverage when the input $x \in \mathcal{X}$ is being adversarially paraphrased, under both senses of split CP and quasi-conditional CP (QCCP).

3.1 Paraphrase-Aware Nonconformity Scores

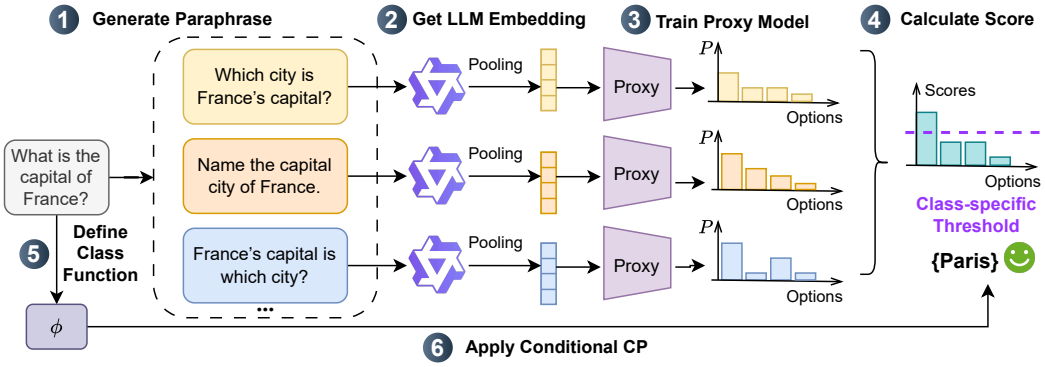

Figure 2: Schematic illustration of our proposed PA score integrated with QCCP.

Method Overview. Our method introduces a *paraphrase-aware* (PA) nonconformity score that captures semantic stability by aggregating predictions across paraphrases. The PA score can be integrated with either standard split CP or QCCP, which provides robustness to adversarially reworded inputs while ensuring statistically valid coverage through the conformal prediction component. At a high level (Figure 2), our approach first generates paraphrases for each question and extracts the corresponding LLM hidden states. A lightweight *proxy classifier* with calibration-aware training then maps these hidden states to confidence probability estimates, from which paraphrase-aware nonconformity scores are derived. Finally, the scores are calibrated with either split CP or QCCP to produce the prediction set for the given question.

Learning a Proxy Classifier for Well-calibrated LLM Uncertainty. Because raw LLM logits may be over-confident and ill-calibrated, we introduce a lightweight proxy classifier $P_\theta(y \mid E(x))$ that approximates the LLM's decision behavior. This proxy takes the mean-pooled last hidden state

of LLM as input and outputs a confidence probability distribution over labels $\mathcal{Y}$. Implemented as a shallow two-layer MLP, it is trained on LLM embeddings $\{E(X_i)\}$ with labels $\{Y_i\}$ using a soft-binned ECE loss (Karandikar et al., 2021) as calibration loss in addition to task loss (cross-entropy for classification task). This design mitigates the poor calibration of raw LLM probabilities while remaining computationally efficient. When the proxy is well calibrated, its outputs are also easy to interpret. For example, if it assigns $P_\theta(\texttt{Paris} \mid E(x)) = 0.8$, this means that roughly 80% of question paraphrases with similar embeddings have "Paris" as the correct answer.

Three Paraphrase-aware Nonconformity Scores. We now turn to the construction of nonconformity scores, the core component of our method. A standard choice in CP is the probability-based score (Sadinle et al., 2019)

$$S_{\text{prob}}(x, y) \;=\; 1 - P_\theta(y \mid E(x)),$$

which simply measures the lack of confidence in assigning label y to question x under the proxy model P_ϕ. While this baseline captures predictive uncertainty for a single prompt, it neglects the instability that often arises when the same question is rephrased. To address this limitation, we define three paraphrase-aware nonconformity scores. From the previous paraphrasing step, we collect a paraphrase set $\mathcal{B}(x) = \{x'_1, \ldots, x'_m\}$ for each question x. Based on this set, the first score

$$S_{\text{mean}}(x, y) = \frac{1}{m} \sum_{x' \in \mathcal{B}(x)} S_{\text{prob}}(x', y),$$

averages the base scores across paraphrases, capturing overall semantic stability. The second,

$$S_{\text{weighted}}(x, y) = \frac{\sum_{x' \in \mathcal{B}(x)} w(x, x')\, S_{\text{prob}}(x', y)}{\sum_{x' \in \mathcal{B}(x)} w(x, x')}, \quad w(x, x') = \exp\big(-\text{sim}(E(x), E(x'))\big),$$

assigns greater weight to paraphrases that are closer in the embedding space, thereby emphasizing local semantic similarity. The third,

$$S_{\text{worst}}(x, y) = \max_{x' \in \mathcal{B}(x)} S_{\text{prob}}(x', y)$$

takes the maximum score across paraphrases, providing a conservative and adversarially robust measure that is sensitive to the hardest rephrasing.

Mean, weighted, and worst-case scores represent a spectrum from efficiency to robustness. Mean yields the most compact sets but may miss rare paraphrases. Weighted emphasizes closer variants to balance compactness and robustness. Worst-case is conservative, guarding against adversarial rephrasings. We compare all three and quantify their trade-offs in Figure 7.

3.2 Split Conformal and Quasi-Conditional Conformal Calibration

To ensure statistically valid coverage, we apply either split CP or its refinement, quasi-conditional CP (QCCP) (Gibbs et al., 2025), on top of our PA nonconformity scores.

Split CP. Given an i.i.d. calibration set $\{(X_i, Y_i)\}_{i=1}^n$, split CP constructs the prediction set for a new input X_{n+1} as

$$\widehat{C}(X_{n+1}) = \{y : S(X_{n+1}, y) \le \widehat{\gamma}_\alpha\},$$

where S is a nonconformity score (cf. Section 3.1) and $\widehat{\gamma}_\alpha$ is the empirical $(1-\alpha)$ quantile of $\{S(X_i, Y_i)\}_{i=1}^n \cup \{\infty\}$. If the distribution of each $S(X_i, Y_i)$ is continuous, Vovk et al. (2005) show that split CP achieves marginal coverage:

$$1 - \alpha \;\le\; \mathbb{P}(Y_{n+1} \in \widehat{C}(X_{n+1})) \;\le\; 1 - \alpha + \frac{1}{n+1}. \tag{1}$$

Split CP is simple and distribution-free, but it relies on a single global threshold $\widehat{\gamma}_\alpha$ for all new inputs X_{n+1}, which can be overly conservative in heterogeneous regions of the input space.

Quasi-conditional CP (QCCP). QCCP refines split CP by adapting conformity thresholds to inputs. In doing so, it achieves guarantees that interpolate between marginal and conditional coverage. More specifically, given a precollected i.i.d. calibration set $\{(X_i, Y_i)\}_{i=1}^n$, it first computes scores

$S_i = S(X_i, Y_i)$ for each $i = 1, \ldots, n$. These scores are then used to fit the following augmented quantile regressor,

$$\widehat{g}_S \in \arg\min_{g \in \mathcal{F}} \frac{1}{n+1} \Big[\sum_{i=1}^{n} \ell_\alpha\big(g(X_i), S_i\big) + \ell_\alpha\big(g(X_{n+1}), S\big) \Big], \tag{2}$$

where $\mathcal{F} = \{\phi(\cdot)^\top \beta : \beta \in \mathbb{R}^d\}$ denotes the linear class over the class mapping ϕ (which we will specify later) and ℓ_α is the pinball loss, defined by $\ell_\alpha(\theta, S) = (S - \theta)\,(\mathbf{1}_{\{S \geq \theta\}} - \alpha)$. At inference time, for a new pair (X_{n+1}, Y_{n+1}) produced by the LLM, we construct the prediction set

$$\widehat{C}(X_{n+1}) = \big\{ y \,:\, S(X_{n+1}, y) \leq \widehat{g}_{S(X_{n+1}, y)}(X_{n+1}) \big\}. \tag{3}$$

The advantage of QCCP lies in the introduction of $\widehat{g}_S$, which adaptively estimates thresholds based on the class function ϕ, rather than relying on the fixed global threshold $\widehat{\gamma}_\alpha$ used in split CP. Consequently, QCCP provides stronger coverage guarantees than the marginal coverage by split CP (see Appendix A for details).

Choice of the Class Function ϕ. We now specify the choice of the class function ϕ. At a high level, we want to group questions by their semantic type (e.g., commonsense vs. factual), so that QCCP can calibrate thresholds within semantically coherent classes and avoid the conservativeness of a single global threshold.

However, in practice, most benchmarks lack explicit type annotations, so we construct them automatically. Specifically, we embed each question–context pair using a pretrained SBERT encoder (`all-MiniLM-L6-v2`) (Reimers & Gurevych, 2019) to obtain sentence embeddings, and then perform K-means clustering on the embeddings of the calibration and test splits. Each sample is assigned to the nearest centroid, and the resulting cluster assignments serve as $\phi(x)$. Mathematically, this amounts to partitioning $\mathcal{X}$ into groups $\{\mathcal{G}_j\}_{j=1}^m$ (with m the total number of clusters) and setting $\phi(x) = \sum_{j=1}^m \mathbf{1}\{x \in \mathcal{G}_j\} \cdot \phi_j$, where ϕ_j denotes the representative vector of cluster j. The hyperparameter K can either be fixed or automatically selected via the silhouette score on calibration embeddings. In short, our choice of ϕ uses sentence embeddings to generate semantically coherent clusters and allows QCCP to condition on latent question classes without requiring manual labels.[1]

Complete Pipeline. We now summarize our full procedure in Algorithm 1. The algorithm integrates paraphrase generation, semantic embeddings, proxy-based uncertainty estimation, and quasi-conditional calibration into a single framework that produces semantically robust prediction sets.

Algorithm 1: Paraphrase-Robust Quasi-Conditional CP

Input: Calibration set $\{(X_i, Y_i)\}_{i=1}^n$, paraphrase generator, embedding map E, proxy model P_θ, class function ϕ, confidence level α, score S (mean, weighted, or worst-case).
Train proxy: Fit P_θ on $\{(E(X_i), Y_i)\}$ using a calibration-aware loss.
Compute scores: For each (X_i, Y_i), compute $S_i \leftarrow S(X_i, Y_i)$ and record class $\phi(X_i)$.
for each test question x **do**
  Initialize $\widehat{C}(x) \leftarrow \emptyset$.
  Generate a set of paraphrases $\mathcal{B}(x)$ and compute embeddings $\{E(x')\}_{x' \in \mathcal{B}(x)}$.
  Determine class label $z \leftarrow \phi(x)$.
  for $y \in \mathcal{Y}$ **do**
    Compute paraphrase-aware score $S(x, y)$.
    Query the class-conditional quantile function $\widehat{g}_z$ from QCCP.
    Add y to $\widehat{C}(x)$ if $S(x, y) \leq \widehat{g}_z(x)$.
  return $\widehat{C}(x)$.

4 Empirical Performance

Datasets. We evaluate our framework on seven benchmark datasets. Five of them are general multiple-choice question answering (MCQA) datasets from the LLM-Uncertainty-Benchmark (Ye

[1] For more implementation details see Appendix D.

et al., 2024), including MMLU, CosmosQA, HellaSwag, HaluDial, and HaluSum. They correspond to different tasks: question answering (**QA**), reading comprehension (**RC**), commonsense inference (**CI**), dialogue response selection (**DRS**), and document summarization (**DS**). Each dataset contains 10,000 MCQA questions. In addition, we include two medical MCQA datasets, **MedMCQA** (Pal et al., 2022) and **MedQA** (Jin et al., 2021), and sample 10,000 questions from each.

Baselines and Evaluation Metrics. We compare our method with two strong nonconformity score baselines: Least Ambiguous set-valued Classifiers (LAC) (Sadinle et al., 2019) and Adaptive Prediction Sets (APS) (Romano et al., 2020). Data are split into training, calibration, and test sets with a 40/30/30 ratio. Evaluation on the test set uses two metrics: **Coverage Rate (CR)**, the fraction of examples where the true label is included in the prediction set, and **Set Size (SS)**, the average number of labels in the prediction set. Unless otherwise specified, the target coverage is $1 - \alpha = 0.90$.

Adversarial Paraphrase Generation. To stress-test our PA score, we generate adversarial paraphrases by prompting a local LLM to rephrase each question while preserving its semantics. We introduce distributional shifts through stochastic generation (temperature sampling) and post-process outputs to ensure they remain valid, well-formed questions. When the model fails to produce a valid paraphrase, we retry with a stricter template or fall back to a simple rule-based rewrite. Additional implementation details are provided in Appendix B.

4.1 PA is Robust to Adversarial Paraphrases across Datasets and CP methods

We first evaluate whether our proposed PA score remains robust under adversarial paraphrasing, where each test question is automatically paraphrased using the adversarial-generated paraphrases described above. On `Qwen2.5-7B-Instruct` (Qwen et al., 2025), we apply different scores to both split CP and QCCP across five MCQA benchmarks (Table 1). We observe that the PA score maintains coverage tightly around the 90% target on both normal and adversarial settings, while producing compact prediction sets. By contrast, APS consistently overshoots coverage (95–97%), indicating unreliable guarantees, and LAC, although close to nominal coverage, exhibits large set-size inflation (from 3.44 to 4.79 under the adversarial setting). These results show that our PA score has two advantages: robustness to distribution shifts and compact prediction sets.

Table 1: Strong performance of the proposed PA nonconformity scores under split CP and QCCP with `Qwen2.5-7B-Instruct` . Results are reported on normal and adversarially (Adv.) paraphrased test sets across five benchmarks. Bold numbers indicate the best.

	QA		RC		CI		DRS		DS	
Method	Normal	Adv.	Normal	Adv.	Normal	Adv.	Normal	Adv.	Normal	Adv.
Coverage Rate (CR, %) Better if closer to 90% *— Split CP*										
LAC	**89.63**	88.43	89.00	**90.33**	91.80	91.27	**89.80**	**90.37**	**89.23**	**90.43**
APS	97.97	99.03	93.00	92.27	95.60	91.70	99.07	90.90	92.30	91.87
PA	**89.63**	**89.23**	**89.77**	90.67	**91.10**	**90.57**	90.77	90.77	91.77	90.93
Coverage Rate (CR, %) Better if closer to 90% *— QCCP*										
LAC	**90.07**	88.47	88.87	90.63	91.63	91.00	95.73	**90.13**	**89.70**	**90.13**
APS	96.10	**90.20**	93.00	92.50	98.87	92.17	92.67	91.40	97.93	91.67
PA	89.67	89.47	**90.10**	**90.33**	**91.23**	**90.33**	**90.87**	91.00	91.90	91.13
Set Size (SS) $\downarrow$ *— Split CP*										
LAC	3.13	4.79	3.41	3.87	1.94	4.18	3.43	4.51	2.42	3.47
APS	5.45	5.92	4.20	4.19	3.44	4.27	5.78	4.51	3.34	3.82
PA	**2.25**	**2.71**	**1.19**	**1.32**	**1.53**	**1.97**	**1.56**	**2.12**	**1.00**	**1.47**
Set Size (SS) $\downarrow$ *— QCCP*										
LAC	3.12	4.76	3.40	3.91	2.04	4.10	4.96	4.49	2.51	3.44
APS	5.05	5.19	4.20	4.22	5.40	4.32	4.22	4.59	5.20	3.82
PA	**2.25**	**2.71**	**1.21**	**1.30**	**1.65**	**1.97**	**1.58**	**2.13**	**1.00**	**1.49**

4.2 Generalization Across LLMs

Next, we evaluate cross-model generalization using two models, `Llama-3.1-8B-Instruct` (Dubey et al., 2024) and `Phi-3-small-8k-Instruct` (Abdin et al., 2024), on MMLU under

both normal and adversarial paraphrase settings. The results mirror the patterns observed in Section 4.1: APS tends to inflate coverage, often overshooting the nominal 90% level, while LAC achieves target coverage but suffers from inflated set sizes under paraphrasing. In contrast, our score (PA) consistently produces the most compact prediction sets and aligns well with the nominal coverage.

Table 2: Cross-LLM generalization on MMLU. Coverage Rate (CR, %) closer to 90% is better; Set Size (SS) $\downarrow$ is better. Results are shown for normal vs. adversarial paraphrase inputs (Adv.).

		`Llama-3.1-8B`				`Phi-3-small`			
		CR (%)		SS		CR (%)		SS	
	Method	Normal	Adv.	Normal	Adv.	Normal	Adv.	Normal	Adv.
Split CP	LAC	90.97	**89.97**	3.65	5.09	98.53	98.67	5.54	5.74
	APS	93.87	100.0	3.95	6.00	98.13	97.97	5.51	5.67
	PA	**89.23**	88.53	**2.51**	**2.81**	**89.27**	**89.70**	**1.97**	**2.59**
QCCP	LAC	91.10	**90.30**	3.67	5.12	95.87	88.47	4.64	4.28
	APS	93.87	99.97	3.96	6.00	91.03	75.83	3.96	2.76
	PA	**89.03**	88.97	**2.51**	**2.85**	**89.23**	**90.00**	**1.96**	**2.61**

4.3 Robustness on Medical MCQA Datasets

Finally, we evaluate whether paraphrase-aware conformal prediction generalizes to domain-specific settings by testing on two medical MCQA benchmarks, MedMCQA (Pal et al., 2022) and MedQA (Jin et al., 2021), using `Qwen2.5-7B-Instruct` under both normal and adversarial paraphrase conditions. As shown in Table 3, the results mirror the patterns observed in the general domain: APS variants often overshoot or fluctuate around the 90% target (e.g., 99.6% on MedQA), while LAC stays closer to nominal coverage but produces inflated sets when paraphrases are introduced. In contrast, our PA score consistently maintains coverage near the target and yields the most compact prediction sets (e.g., 2.6–3.1 on MedMCQA vs. 4.4–4.6 for LAC/APS). These findings confirm that PA robustly preserves both coverage and efficiency even in the medical domain, demonstrating its ability to generalize beyond general-purpose benchmarks.

Table 3: Results on two medical QA benchmarks with `Qwen2.5-7B-Instruct`. Coverage rate (CR, %) closer to the nominal 90% is better, while smaller set size (SS, $\downarrow$) denotes better.

		MedMCQA				**MedQA**			
		CR (%)		SS		CR (%)		SS	
	Method	Normal	Adv.	Normal	Adv.	Normal	Adv.	Normal	Adv.
Split CP	LAC	89.43	**89.50**	4.46	4.48	91.67	**89.43**	4.47	5.00
	APS	91.30	91.67	4.47	4.45	99.60	88.80	5.85	4.72
	PA	**89.67**	91.47	**2.59**	**3.01**	**89.73**	89.13	**4.02**	**4.31**
QCCP	LAC	**90.03**	**90.43**	4.55	4.60	91.23	89.87	4.64	5.02
	APS	92.03	91.50	4.59	4.52	93.67	89.57	4.92	4.81
	PA	90.43	91.00	**2.77**	**3.09**	**89.90**	**89.90**	**4.03**	**4.39**

5 In-Depth Analysis and Ablation Studies

Proxy Model Improves Accuracy and Calibration. In our work, the proxy model P_θ is a lightweight two-layer MLP trained on top of frozen hidden states from the LLM. We argue that P_θ yields more accurate and better-calibrated distributions than raw LLM logits. As shown in Figure 3, this proxy consistently outperforms logits across benchmarks on all calibration metrics. For example, accuracy improves substantially (**CI** $0.263 \rightarrow 0.728$, **QA** $0.287 \rightarrow 0.551$), while both Brier score and NLL decrease (**RC** Brier $0.126 \rightarrow 0.050$, NLL $1.592 \rightarrow 0.647$). Here, Brier captures the mean squared error between predicted probabilities and true one-hot labels, while NLL penalizes models that assign low probability to the correct answer, both measuring the calibration quality. This improvement arises because raw logits are optimized for next-token prediction rather than calibrated

posteriors, which leads to miscalibration and poor class separation. In contrast, the proxy leverages hidden states, which encode richer task-relevant signals, and is trained with a calibration-aware loss. As a result, it has better accuracy and calibration.

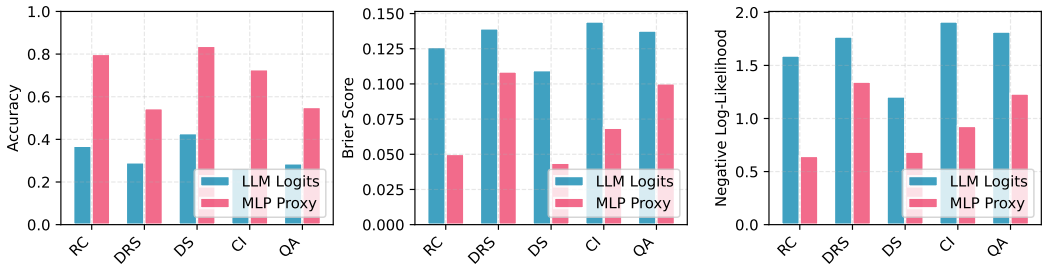

Figure 3: The proxy improves both task performance and calibration consistently. Evaluation of prediction accuracy and calibration metric using proxy model vs. raw LLM logits across datasets. Left: accuracy (higher is better). Middle: Brier (lower is better). Right: NLL (lower is better).

Proxy Model Decreases Prediction Set Size. Next, we examine the effect of the proxy model on prediction set size. To this end, we ablate the proxy model by computing the PA score directly from LLM logits. Across all five datasets, coverage remains close to the nominal 90%, but *set sizes increase substantially* without the proxy under both split CP (in Appendix, Figure 8) and QCCP (in Figure 4). This shows that the proxy produces better-calibrated predictive distributions, which translate into materially smaller sets at comparable coverage.

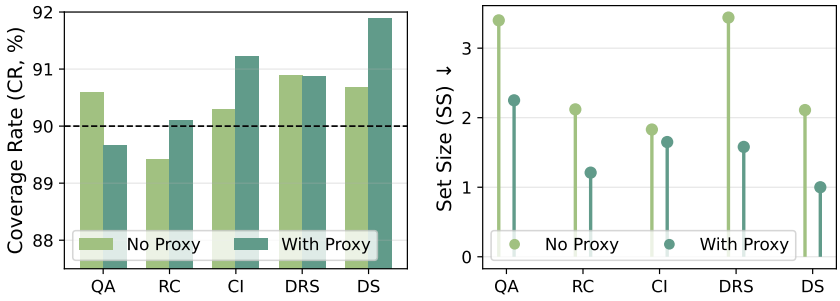

Figure 4: Effect of removing the proxy model under the quasi-conditional CP setting. Coverage remains close to 90% (dashed line), but set size increases notably without the proxy model.

Better Class-conditional Coverage. We group questions into semantic classes (or clusters) and evaluate class-conditional coverage by visualizing empirical coverage rates on different classes. As shown in Figure 5, PA consistently outperforms LAC and APS across different classes.

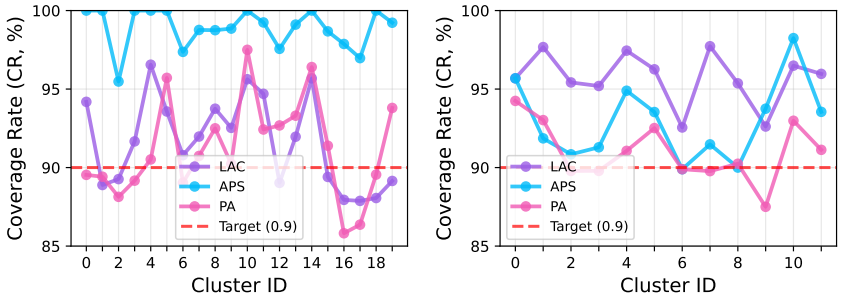

Figure 5: Evaluation of class-conditional coverage with QCCP on the HellaSwag dataset (left) and HaluDial dataset (right). Results on additional datasets are provided in Figure 9.

Effect of Coverage Levels. To examine the effect of the level α, we vary $\alpha \in \{0.2, 0.05, 0.01\}$ and evaluate both coverage and prediction set size again. As shown in Figure 6, our PA score consistently tracks the nominal $(1 - \alpha)$ target while maintaining compact sets. For example, at $\alpha = 0.05$, it achieves $\approx 95\%$ coverage with an average set size of 2.7, compared to APS's inflated 5.0. In contrast, APS again overshoots, while LAC produces smaller but unstable sets. Overall, PA consistently outperforms across different levels.

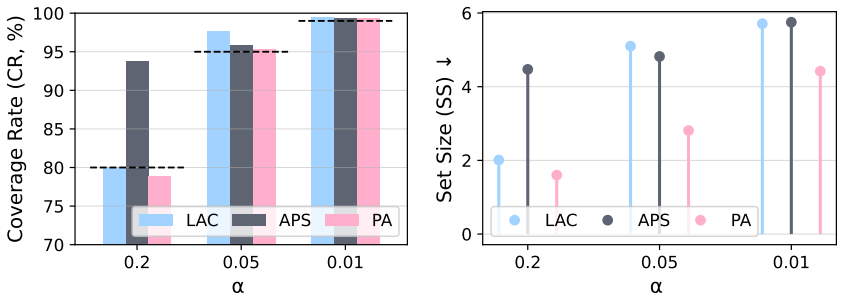

Figure 6: Coverage rate (left) and prediction set size (right) vs. α on MMLU under QCCP.

Larger Paraphrase Budgets Yield Smaller Prediction Sets. We now study the effect of the paraphrase budget m for `Qwen2.5-7B-Instruct` (Qwen et al., 2025) on the MMLU dataset. As shown in Table 4, increasing $m \in \{2, 4, 6\}$ leads our PA score to achieve higher coverage and smaller set sizes. This is intuitive: more paraphrases provide richer semantic views of each question, which reduces variance in the aggregated score and enables more precise calibration.

Table 4: Effect of paraphrase budget m.

	m=2	m=4	m=6
Coverage Rate (CR, %)			
PA (marginal)	89.10	89.47	**89.63**
PA (conditional)	**89.87**	89.60	89.67
Set Size (SS)			
PA (marginal)	2.42	2.32	**2.25**
PA (conditional)	2.49	2.34	**2.25**

Comparison of Different Scores. Finally, we compare the three PA scores, namely Mean, Weighted, and Worst, introduced in Section 3.1. We evaluate them under both split CP and QCCP settings. As shown in Figure 7, the Mean score consistently yields the most compact sets (whose SS ≈1.5–1.6) while staying close to the target 90% coverage. The Weighted variant produces moderately larger sets (whose SS ≈2.3) without noticeable gains in coverage, whereas the Worst variant greatly inflates set size (whose SS ≈3.0) and overshoots coverage (which ≥95%). QCCP mitigates some of the overshoot for Mean and Weighted but largely preserves their relative ranking. Overall, Mean offers the best efficiency, Weighted provides only a mild trade-off, and Worst proves overly conservative.

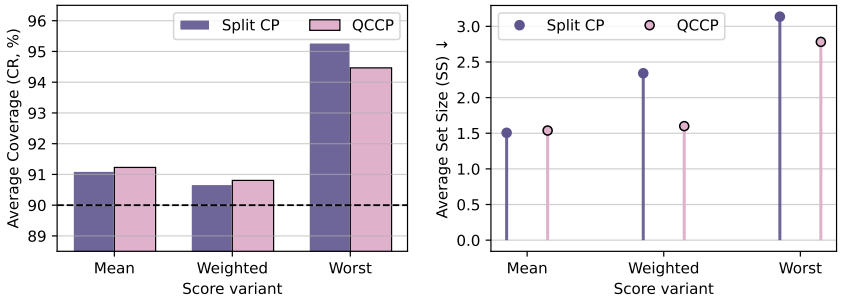

Figure 7: Comparison of different scores. Left: average coverage (closer to 90% is better). Right: average set size (smaller is better).

6 Conclusion

In this work, we introduced a framework for paraphrase-robust conformal prediction by designing paraphrase-aware nonconformity scores and applying them to both split CP and quasi-conditional CP. Our method preserves theoretical coverage guarantees while yielding substantially smaller prediction sets than logit-based baselines. Experiments on five general QA and two medical QA benchmarks demonstrate that it remains reliable under adversarial paraphrasing and generalizes across model families. More broadly, our work illustrates how CP can be adapted to address semantic invariance and distribution shifts in LLM uncertainty quantification. Promising directions include extending paraphrase-robust scores to free-form generation, integrating them with selective abstention policies, and exploring theoretical bounds under broader perturbation classes.

Ethical Statement

All authors have read and adhere to the conference Code of Ethics. We acknowledge the use of large language models (LLMs) for limited purposes in this paper, only for polishing the writing and assisting with literature search. All LLM-generated content was carefully reviewed and verified by the authors, who take full responsibility for the final manuscript.

Reproducibility Statement

To support reproducibility, we provide an anonymized GitHub repository link at the end of the abstract containing our codebase. Detailed descriptions of the dataset are included in Appendix C, and all hyperparameter settings are reported in Appendix E.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Oleksandr Balabanov and Hampus Linander. Uncertainty quantification in fine-tuned LLMs using loRA ensembles. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*, 2025. URL https://openreview.net/forum?id=L8KaEqT70q.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *ICML*, 2024. URL https://openreview.net/forum?id=rJVjQSQ8ye.

Dylan Bouchard and Mohit Singh Chauhan. Uncertainty quantification for language models: A suite of black-box, white-box, llm judge, and ensemble scorers. *arXiv preprint arXiv:2504.19254*, 2025.

John Cherian, Isaac Gibbs, and Emmanuel Candes. Large language model validity via enhanced conformal prediction methods. In *Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=JD3NYpeQ3R.

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *AAAI Conference on Artificial Intelligence*, 2025.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in llms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Jinhao Duan, Hao Cheng, Shiqi Wang, et al. Shifting attention to relevance: Towards predictive uncertainty quantification of free-form large language models. In *ACL*, 2024. URL https://aclanthology.org/2024.acl-long.276/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Yu Feng, Phu Mon Htut, Qi Zheng, et al. Diverseagententropy: Quantifying black-box llm uncertainty through diverse perspectives and multi-agent interaction. *arXiv preprint arXiv:2410.08174*, 2024. URL https://openreview.net/forum?id=AJAStQYZaL.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2336–2346, 2024.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. qkaf008, 2025.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YzyCEJlV9Z.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243. URL https://aclanthology.org/D19-1243/.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 284–312, 2024a.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. Calibrating long-form generations from large language models. In *Findings of EMNLP*, 2024b.

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

Archit Karandikar, Nicholas Cain, Dustin Tran, Balaji Lakshminarayanan, Jonathon Shlens, Michael C Mozer, and Becca Roelofs. Soft calibration objectives for neural networks. *Advances in Neural Information Processing Systems*, 34:29768–29779, 2021.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024. URL https://arxiv.org/abs/2406.15927.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. 2023. URL https://arxiv.org/abs/2305.18404.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397/.

L López, Shaza Elsharief, Dhiyaa Al Jorf, Firas Darwish, Congbo Ma, and Farah E Shamout. Uncertainty quantification for machine learning in healthcare: A survey. *arXiv preprint arXiv:2505.02874*, 2025.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9329–9345, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.627/.

Zachary Nado, Neil Band, Mark Collier, Alexander D'Amour, Josip Djolonga, Sebastian Farquhar, Andrew Foong, Alex Kendall, Andrey Malinin, Daniel Muñoz, et al. Uncertainty baselines: benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2022. URL https://arxiv.org/abs/2106.04015.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LOH6qzI7T6.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=pzUhfQ74c5.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=pzUhfQ74c5.

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

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33:3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3), September 2025. ISSN 0360-0300. doi: 10.1145/3744238. URL https://doi.org/10.1145/3744238.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. Re-examining calibration: The case of question answering. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2814–2829, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.204. URL https://aclanthology.org/2022.findings-emnlp.204/.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. API is enough: Conformal prediction for large language models without logit-access. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 979–995, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.54. URL https://aclanthology.org/2024.findings-emnlp.54/.

Sina Tayebati, Divake Kumar, Nastaran Darabi, et al. Learning conformal abstention policies for adaptive risk management in large language and vision-language models. *arXiv preprint arXiv:2502.06884*, 2025. URL https://arxiv.org/abs/2502.06884.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. Exploring predictive uncertainty and calibration in nlp: A study on the impact of method & data scarcity. In *Findings of EMNLP*, 2022. URL https://aclanthology.org/2022.findings-emnlp.198/.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. Calibrating large language models using their generations only. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15440–15459, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.824. URL https://aclanthology.org/2024.acl-long.824/.

Harit Vishwakarma, Alan Mishler, Thomas Cook, Niccolo Dalmasso, Natraj Raman, and Sumitra Ganesh. Prune 'n predict: Optimizing LLM decision-making with conformal prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=5g6LPR0Dlx.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

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

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris N. Metaxas, and Hao Wang. BLob: Bayesian low-rank adaptation by backpropagation for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL https://openreview.net/forum?id=MaDykgj4Ru.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. ConU: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6886–6898, 2024b.

Zhiyuan Wang, Jinhao Duan, Qingni Wang, et al. Coin: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. *arXiv preprint arXiv:2506.20178*, 2025a. URL https://arxiv.org/abs/2506.20178.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Sconu: Selective conformal uncertainty in large language models, 2025b. URL https://arxiv.org/abs/2504.14154.

Guang Yang and Xinyang Liu. Conformal sets in multiple-choice question answering under black-box settings with provable coverage guarantees. *arXiv preprint arXiv:2508.05544*, 2025. URL https://arxiv.org/abs/2508.05544.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking LLMs via uncertainty quantification. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=L0oSfTroNE.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

A Theoretical Guarantee of Quasi-Conditional CP

Lemma A.1 (Theorem 2 in (Gibbs et al., 2025)) *Assume* $(X_i, Y_i)_{i=1}^{n+1}$ *are i.i.d. and that* $S(X, Y) \mid X$ *has a continuous distribution. Then, for any* $f \in \mathcal{F}$*,*

$$\left| \mathbb{E}\Big[f(X_{n+1}) \left(\mathbf{1}\{Y_{n+1} \in \widehat{C}(X_{n+1})\} - (1-\alpha) \right) \Big] \right| = O\left(\frac{1}{n+1} \right).$$

Lemma A.1 establishes the quasi-conditional coverage guarantee for QCCP. This guarantee is strictly stronger than the marginal coverage of split conformal prediction (the case where $\mathcal{F}$ contains only constant functions) and strictly weaker than full conditional coverage (which would require $\mathcal{F}$ to be all measurable functions), providing a principled middle ground. In practice, computing equation 3 reduces to a convex optimization; see Section 4 of Gibbs et al. (2025) for details.

B Adversarial Paraphrase Generation

We only use this procedure to generate adversarial paraphrases for the input question used in the adversarial setting evaluation. The set of adversarially reworded questions is not directly used for calculating the PA score, but is used as another set of input questions.

Adversarial Prompt Pool. We maintain 7 short templates (e.g., "Rephrase this question using varied vocabulary and phrasing: {question}"). One template is sampled per input to induce lexical or syntactic variety without changing semantics.

Batched Generation. Given a batch of n questions, we form n prompts and generate once with: `temperature`=0.7, `top_p`=0.9, `do_sample`=True, `eos/pad_token_id` aligned. We decode with `skip_special_tokens=True`.

Cleaning & Validation. We remove the prompt prefix and extract the first question sentence via regex matching the earliest "?". We then strip boilerplate ("Rephrase:", "Paraphrase:", "Here's a ..."), quotes/bullets/code blocks, and normalize whitespace. The candidate paraphrase must satisfy: (i) non-empty, (ii) case-insensitive $\neq q$, (iii) ends with "?". Duplicates within the batch are dropped.

Retry & Fallback. Failures are retried with a stricter template: single line, no preface, $\leq$20 words, must end with "?". Remaining failures are rewritten by a deterministic rule set that preserves meaning (e.g., "Which of the following"$\rightarrow$"Which option", add trailing "?", etc.).

Algorithm 2: Adversarial Paraphrase Pipeline (per question q)

1: Sample a paraphrase template; compose prompt $p(q)$.
2: Generate with (T=0.7, top-p=0.9, sampling).
3: Decode; strip prompt prefix; take first sentence ending with "?".
4: Clean and validate; if valid, return $\tilde{q}$.
5: Else: retry with strict one-line template; clean and validate.
6: Else: apply rule-based fallback; return $\tilde{q}$.

C Dataset details

We used five general and two medical MCQA datasets to show that our paraphrase-aware score can be broadly used and can adapt to high-stakes scenarios where calibration is important. The five general datasets test a wide range of LLM capabilities and have been used in previous CP benchmarking papers (Ye et al., 2024; Vishwakarma et al., 2025). We use the five general datasets processed by Ye et al. (2024). Each of the datasets contains 10,000 questions.

QA, MMLU (Hendrycks et al., 2021): MMLU is a dataset designed to test the general knowledge and *question-answering* abilities of LLMs. Question topics range from sociology and high school geography to electrical engineering and abstract algebra.

RC, CosmosQA (Huang et al., 2019): CosmosQA focuses on gauging an LLM's *reading comprehension* abilities. The LLM is given a short paragraph and is then asked to answer a follow-up question based on commonsense reasoning.

CI, HellaSwag (Zellers et al., 2019): HellaSwag evaluates if LLMs can use *commonsense inference* to construct a realistic and meaningful continuation of a given scenario. HellaSwag is deliberately designed so that LLMs struggle with questions that humans could normally answer with high confidence.

DRS, HaluEval (Li et al., 2023): HaluEval consists of hallucinated LLM responses to user queries. A subset of HaluEval contains queries that relate to *dialogue response selection*: the LLM must be able to choose a logical response for a conversation. We refer to this part of HaluEval as HaluDial.

DS, HaluEval (Li et al., 2023): The HaluEval dataset also has hallucinated *document summaries*. In an MCQA setting, the LLM must determine which summary in the answer choices is most relevant to a provided document. We refer to this part of HaluEval as HaluSum.

MedMCQA (Pal et al., 2022): MedMCQA is a large-scale multiple-choice medical QA dataset comprising over 194,000 entrance-exam–style questions spanning 2400 healthcare topics and 21 medical subjects. We select a subset of 10,000 single-answer questions from MedMCQA (i.e. where exactly one option is marked correct) for our experiments.

MedQA (Jin et al., 2021): MedQA is a medical exam QA dataset derived from professional medical board exams (e.g. USMLE), providing each question paired with candidate answer options and corresponding references. In our work, we use only the US-part of MedQA, and further restrict to 10,000 multiple-choice items that have exactly one correct answer.

D Detailed method implementation

Paraphrase generation for calculating PA score. Paraphrases were generated by prompting `Qwen2.5-7B` (Qwen et al., 2025) with the following query: "Rephrase the following question in your own words (preserving its meaning): Original question: {question} \n Rephrased question:". Additional details, such as any context or the answer choices, were not included with the original question. A total of 6 paraphrases were generated per question. Paraphrases that were equivalent to the original question or were previously generated were not included in the final set. The temperature of `Qwen2.5-7B` (Qwen et al., 2025) was set to 1.1 to encourage diverse responses.

LLM embeddings. We next obtained the LLM embeddings for each question and paraphrase. These LLM embeddings are used to train the proxy model and serve as the LLM's representation of the query. We treated the paraphrases as new samples in the dataset and assigned them the same answer choices and correct label as their parent question. For each sample, the input prompt included the question and the list of answer choices without the correct label. The LLM embedding was extracted from the final hidden layer; this layer simultaneously encodes the LLM's understanding of the question and its predicted answer. The LLM logits were also retrieved by finding the raw score corresponding to each answer represented as a token (e.g. 'A', 'B', etc.). We then applied softmax to the logits to obtain a probability distribution over the answer choices. For each dataset except the DS dataset, the LLM was provided with 2 example questions and their correct labels. Due to the long context needed for the questions in the DS dataset, only 1 example was provided to the LLM in this case. We followed this procedure to get the embeddings and logits of the instruction-tuned versions of `Qwen2.5-7B` (Qwen et al., 2025) , `Llama-3.1-8B` (Dubey et al., 2024), and `Phi-3-small-8k` (Abdin et al., 2024).

Proxy model. The proxy model calibrates the LLM's probability distribution of the answer choices across the paraphrases. We used a 2-layer MLP with ReLU activation as our proxy model. The input to the proxy model is an LLM's embedding for a question or paraphrase, and the output is a vector with dimension $|\mathcal{Y}|$. The loss function used to train the proxy model is $L_{total} = L_{CE} + \lambda_{ece} \times L_{ECE}$, where L_{CE} is the standard cross-entropy loss for multiclass classification and L_{ECE} is the soft-binned expected calibration error loss (Karandikar et al., 2021). λ_{ECE} is a hyperparameter that can be tuned to increase or decrease the relative importance of calibration in different contexts. Out of the 4,000 questions in the training set, 600 questions were used as the validation set. Since each question has 6 paraphrases, the effective training and validation set sizes are 23,800 and 4,200, respectively. We conducted grid search on the following hyperparameters: learning rate, weight decay, λ_{ECE}, hidden dimension, and the batch size. The hyperparameters used for each dataset and each model are provided in Appendix E.

Score calculation. For our baselines, we only use logits from the *original* question. The LAC score (Sadinle et al., 2019) is defined as $s_{\text{LAC}}(x, y) = 1 - f(x)_y$, where $f(x)_y$ is the softmax probability of label y for a question x. The APS score (Romano et al., 2020) is $s_{\text{APS}}(x, y) = \sum_{y' \in \mathcal{Y}:\, f(x)_{y'} \geq f(x)_y} f(x)_{y'}$, i.e., the cumulative probability of labels ranked at least as high as y. To find the score of a question x using the PA method, the LLM embeddings of its paraphrases (represented by the set $\mathcal{B}(x)$) are inputted into the trained proxy model. Then, the score for answer choice y of question x is given by $S_{\text{mean}}(x, y)$, $S_{\text{weighted}}(x, y)$, or $S_{\text{worst}}(x, y)$ (see Section 3.1). Note that only the scores of the 6 paraphrases are used in the PA formulas. The LLM's embedding of the original question does not contribute to that question's final PA score. This is in contrast to LAC and APS, which only rely on the logits of the original question and do not factor in the paraphrases.

Split conformal prediction. Split CP (Vovk et al., 2005) uses a separate calibration dataset to calculate a global score threshold that determines the prediction sets for the test set. For each calibration example, we use the nonconformity score from the LLM's softmax probabilities or proxy model's probability distribution. These nonconformity scores are collected and the $(1 - \alpha)$ quantile is estimated with a finite-sample correction $q_\alpha = \text{Quantile}(\{s_i\}, \lceil (n+1)(1-\alpha) \rceil / n)$. At test time, each example's prediction set is formed by including all labels whose score is below q_α, with a fallback to the most probable label if the set is empty. This procedure ensures that the empirical coverage approaches the nominal $1 - \alpha$ guarantee.

Quasi-conditional conformal prediction. In contrast to split CP, QCCP (Gibbs et al., 2025) calculates class-specific thresholds for pre-defined classes. A function ϕ is defined that assigns a class to a question based on its features. In the case that ϕ is a constant function, then QCCP is equivalent to split CP; we use this ϕ function (ϕ = intercept) to obtain all split CP results. In our QCCP analysis, we focus on a ϕ function that takes in the embedding of each question generated by `all-MiniLM-L6-V2` (Reimers & Gurevych, 2019). K-means clustering is used to separate the questions into clusters, with each cluster acting as a class, and a different score threshold is defined for each cluster. For a given test question, either a one-hot encoding of its assigned cluster or a vector of its embedding's distances to the three closest cluster centroids (dist3) can be used to calculate the threshold. The coverage rates reported using QCCP are calculated marginally with the exception of Figure 5 and Figure 9, which displays class-wise coverages. We used the `conditionalconformal` package (Gibbs et al., 2025) to implement QCCP.

E HYPERPARAMETER SETTINGS FOR REPRODUCIBILITY

Tables 5 and 6 list the framework hyperparameters for the five general MCQA datasets and the two medical MCQA datasets, respectively. Tables 7 and 8 report the corresponding hyperparameters for the proxy model on the general and medical datasets.

Table 5: Hyperparameters for QA, RC, CI, DRS, DS. We train an MLP proxy with ECE regularization and compute paraphrase-aware scores using 6 paraphrases per question. Evaluation includes QCP with SBERT-clustered Φ and plain CP with intercept Φ, using the same manifest-based splits and α=0.1.

Setting / Hyperparameter	**QA, RC, CI, DRS, DS**
General	
Base LLM (for reps/logits)	Qwen2.5-7B-Instruct
Random seed	42
Samples per question (n)	7 (1 base + 6 paraphrases)
Options per item ($\lvert\mathcal{Y}\rvert$)	6 (A–F)
Proxy model (training)	
Input dim	3584
Architecture	MLP: 3584 $\rightarrow h$ (ReLU) $\rightarrow$ 6
Optimizer	Adam
Max epochs / patience	200 / 20
Loss	CE + soft-binned ECE (15 bins, temp 0.1))
Paraphrase-aware score computation	
Paraphrases per question	6
Metric used	`S_mean`
Conformal prediction / QCP evaluation	
Prompting / ICL	`base` / `icl1`
Error level (α)	0.1
Split config	manifest: `tr0.4 / cf0.3 / tf0.3`
QCP Φ mode	`cluster_sbert`
SBERT model	MiniLM-L6-v2
Cluster selection	auto
Φ representation	dist3
Embedding norm / mini-batch k-means	on / on
Also reported (plain CP)	`intercept` Φ (same splits, α)

Table 6: Hyperparameters for MedMCQA-10k and MedQA-10k. We train an MLP proxy with ECE regularization and compute paraphrase-aware scores using 6 paraphrases per question. Evaluation includes QCP with SBERT-clustered Φ and plain CP with intercept Φ, using the same manifest-based splits and α=0.1.

Setting / Hyperparameter	**MedMCQA-10k**	**MedQA-10k**
General		
Base LLM (for reps/logits)	Qwen2.5-7B-Instruct	
Random seed	42	
Samples per question (n)	7 (1 base + 6 paraphrases)	
Options per item (K)	6 (A–F)	
Proxy model (training)		
Input dim	3584	
Architecture	MLP: 3584 $\rightarrow$ h (ReLU) $\rightarrow$ 6	
Optimizer	Adam	
Max epochs / patience	50 / 5	
Loss	CE + soft-binned ECE (15 bins, temp 0.1)	
Paraphrase-aware score computation		
Paraphrases per question	6	
Metric used	`S_mean`	
Conformal prediction / QCP evaluation		
Prompting / ICL	`task` / `icl1`	
Error level (α)	0.1	
Split config	manifest: `tr0.4 / cf0.3 / tf0.3`	
QCP Φ mode	`cluster_sbert`	
SBERT model	MiniLM-L6-v2	
Cluster selection / K	fixed / 20	
Φ representation	one-hot (cluster ID)	
Embedding norm / mini-batch k-means	off / off	
Also reported (plain CP)	`intercept` Φ (same splits, α)	

Table 7: Proxy model specific hyperparameters for the 5 general datasets

Dataset	**Hidden dimension h**	**Batch size**	**Learning rate**	**Weight decay**	λ_{ECE}
MMLU	256	64	1e-3	0.0	0.5
CosmosQA	512	128	1e-3	0.0001	0.5
HellaSwag	256	128	1e-3	0.0001	0.5
HaluDial	256	64	1e-3	0.0	0.5
HaluSum	256	128	1e-4	0.0	0.5

Table 8: Proxy model specific hyperparameters for MedMCQA and MedQA.

Dataset	**Hidden dimension h**	**Batch size**	**Learning rate**	**Weight decay**	λ_{ECE}
MedMCQA	256	64	1e-3	0.0	0.5
MedQA	256	64	1e-3	0.0	0.5

F Additional Results for Analysis and Ablation Studies

F.1 Proxy Model Ablation under Split CP

We report proxy model ablation results under split CP in Figure 8; the same trend holds under QCCP (Figure 4), where removing the proxy leaves coverage near 90% but substantially enlarges set sizes.

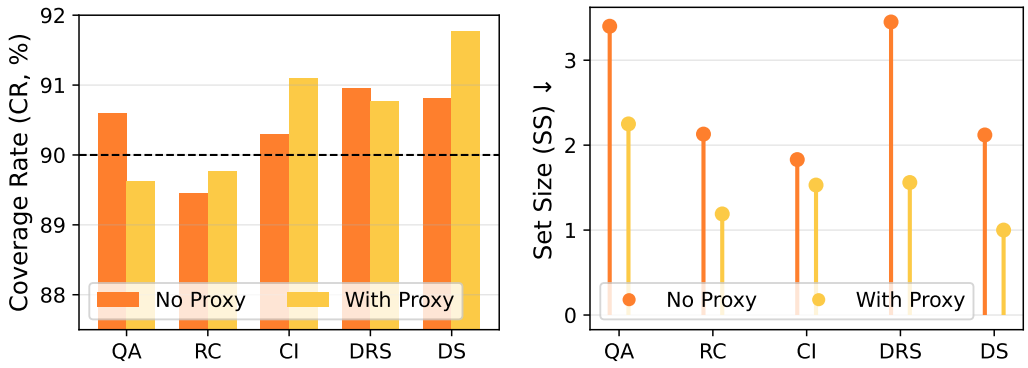

Figure 8: Effect of removing the proxy model under the split CP setting. Coverage remains close to 90% (dashed line), but set size increases notably without the proxy model.

F.2 Class Level Evaluation for Additional Datasets

We report class-conditional coverage for the remaining datasets in Figure 9; the results mirror those in the main text, with PA consistently achieving better class-level coverage than LAC and APS across classes.

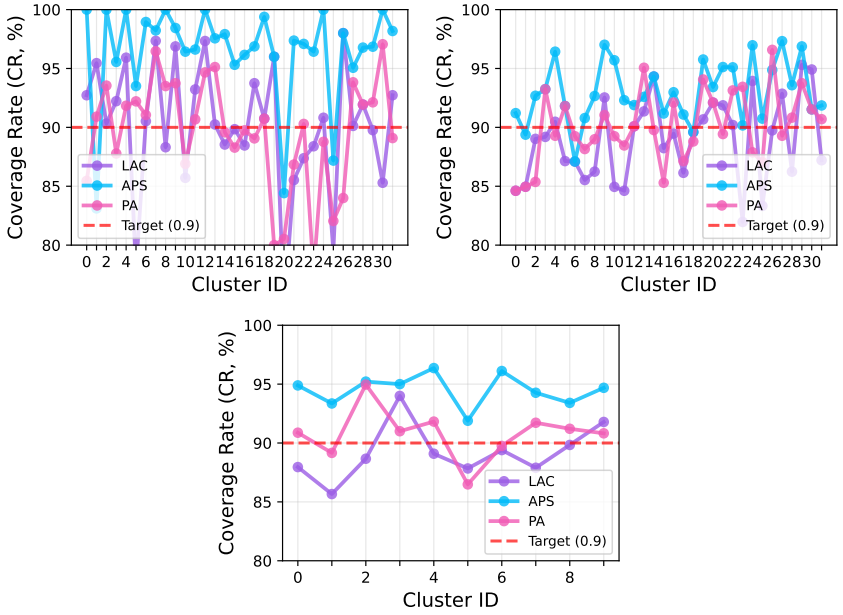

Figure 9: Evaluation of class-conditional coverage with QCCP on the MMLU (top left), CosmosQA, (top right), and HaluSum (bottom center) datasets.

F.3 Different Risk Level Analysis under Split CP

We report results at different user-specified risk levels under split CP in Figure 10. The trends mirror those of QCCP in the main text: PA tracks the nominal $(1-\alpha)$ target more closely while keeping sets compact, whereas APS overshoots and LAC has larger set sizes for low values of α.

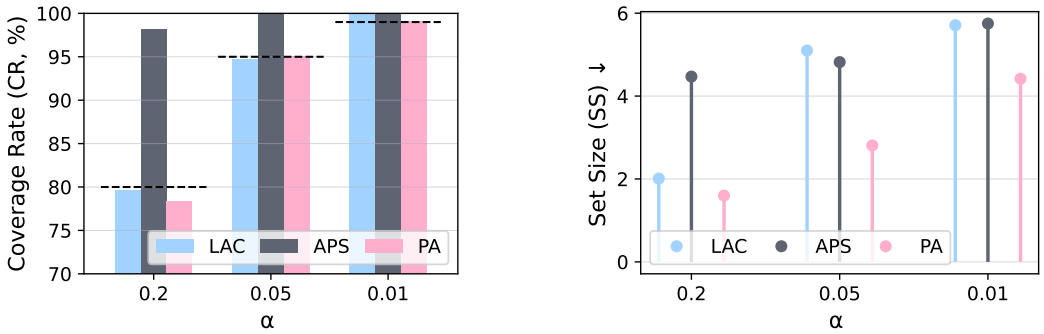

Figure 10: Coverage rate (left) and prediction set size (right) vs. α on MMLU under Split CP.