ICLR2025 Spotlight Paper

Sample then Identify: A General Framework for Risk Control and Assessment in Multimodal Large Language Models

Qingni Wang[1], Tiantian Geng[2,3], Zhiyuan Wang[1], Teng Wang[2,4],
Bo Fu[1*] , Feng Zheng[2*]

[1] School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2] Southern University of Science and Technology
[3] University of Birmingham
[4]The University of Hong Kong

Presented by: Qingni Wang

* Corresponding co-authors

Background

- Multimodal Large Language Models (MLLMs) exhibit promising advancements across various tasks, yet they still encounter significant trustworthiness issues.
- Uncertainty quantification (UQ) can provide valuable insights into the reliability of model responses, facilitating risk control and assessment.
- Split Conformal prediction (SCP) is a distribution-free and model-agnostic approach to UQ, which transforms any heuristic notion of uncertainty into a statistically rigorous one.

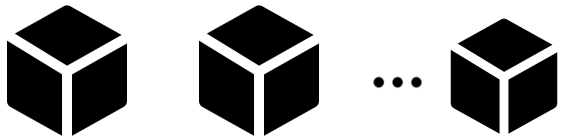
Motivation

Adapting SCP for MLLMs in practical closed-ended and open-domain VideoQA tasks presents several challenges:

- Adaptability. Unline tasks equipped with fixed options covering the correct answers, the output space for each VideoQA task is unbounded and there may not be an acceptable response by sampling multiple generations;
- Reliability. Logits or verbalized confidence levels may be miscalibrated, leading to biased prediction sets;
- Flexibility. For some API-only MLLMs, users lack access to internal model information like sequence logits.

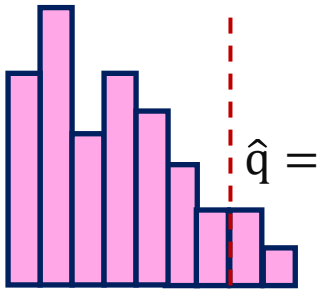
Our Framework TRON

Calibration Set



$$\left(x_1,\ y_1^*,\ \left\{y_j^{(1)}\right\}_{j=1}^{m}\right) \quad \cdots \quad \left(x_n,\ y_n^*,\ \left\{y_j^{(n)}\right\}_{j=1}^{m}\right)$$

$$s\left(x_i,\ y_{ref}^{\{i\}}\right)$$



Non-conformity Scores s_i

risk level

$$\hat{q} = quantile\left(\{s_i\}_{i=1}^{n},\ \frac{\lceil (1-\beta)(n+1)\rceil}{n}\right)$$

Correlate the non-conformity score with the uncertainty state of y_{ref} within the sampling set $\left\{y_j^{(i)}\right\}_{j=1}^{m}$, which is semantically equivalent to $y_{ref}^{\{i\}}$, assuming that we can obtain at least one acceptable response by sampling m generations.

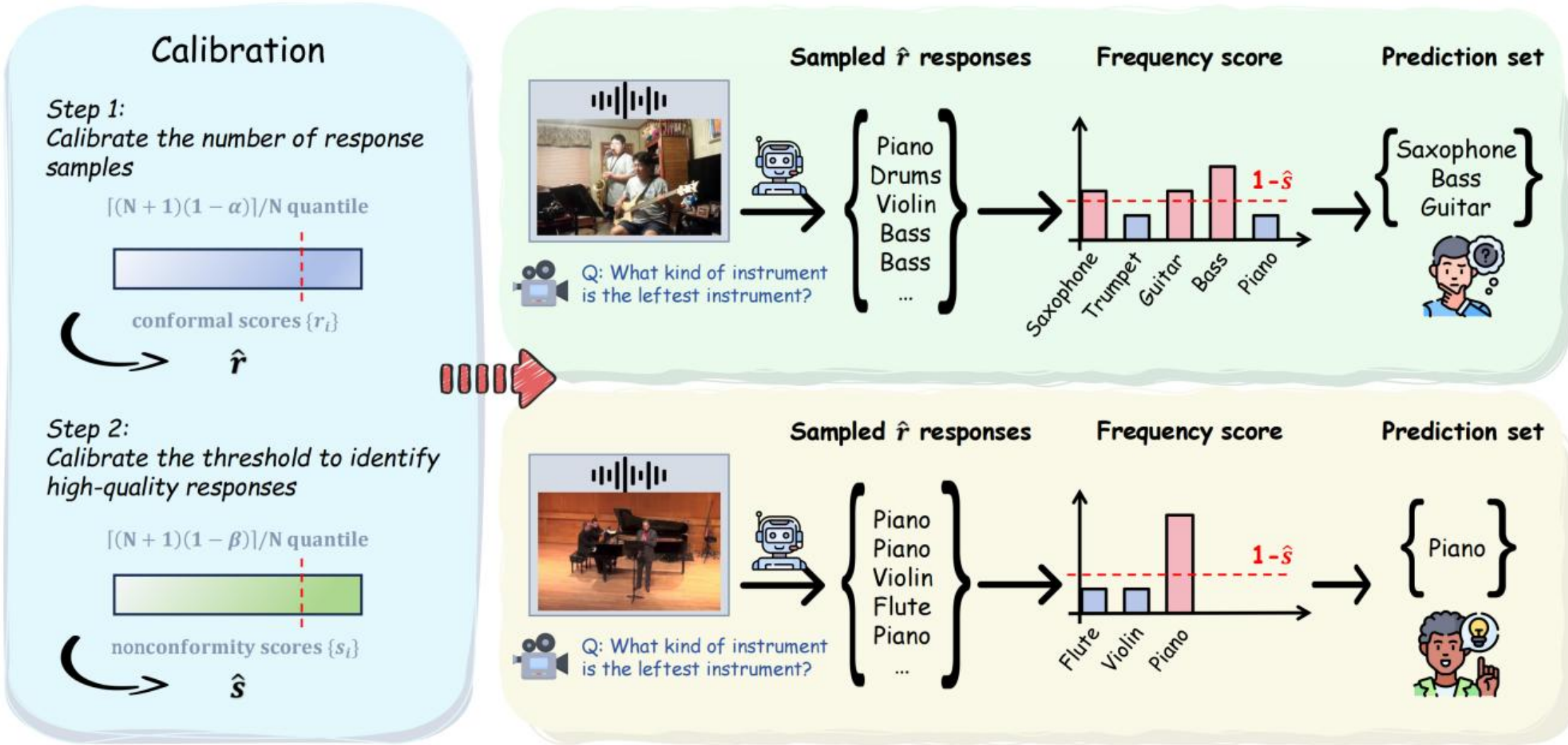
$$\frac{\lceil (N+1)(1-\alpha)\rceil}{N}\ \textbf{quantile}$$

conformal scores $\{r_i\}_{i=1}^{n}$

*Before the identification of reliable generations, we can also manage the error rate of the sampling set failing to cover acceptable responses by developing the **conformal score** that determines the minimum sampling size.*

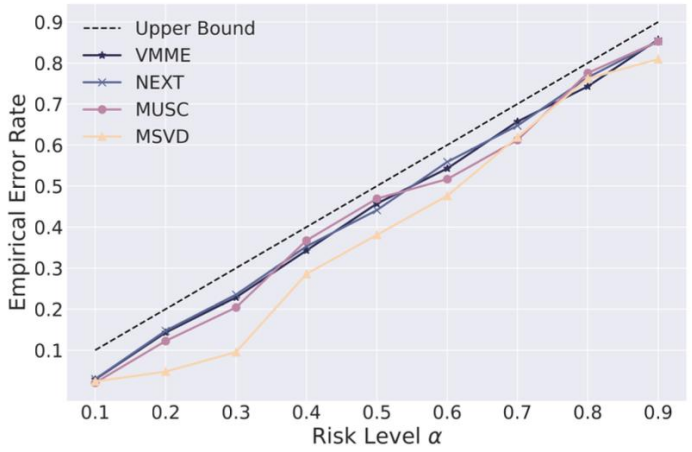
$$r_i = r(x_i,\ y_i^*) := \sup\left\{M_i:\ \forall M_i^{'} < M_i,\ y_i^* \notin \left\{y_j^{(i)}\right\}_{j=1}^{M_i^{'}}\right\}$$

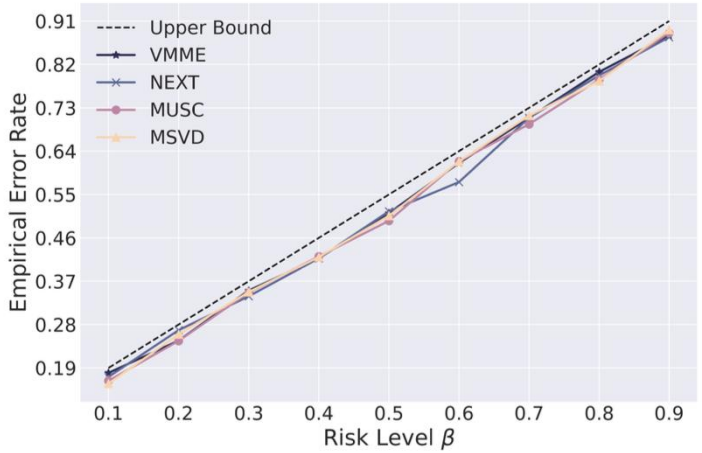
We set the sampling size of each test sample to the $\hat{r}$ based on the other risk level of α.



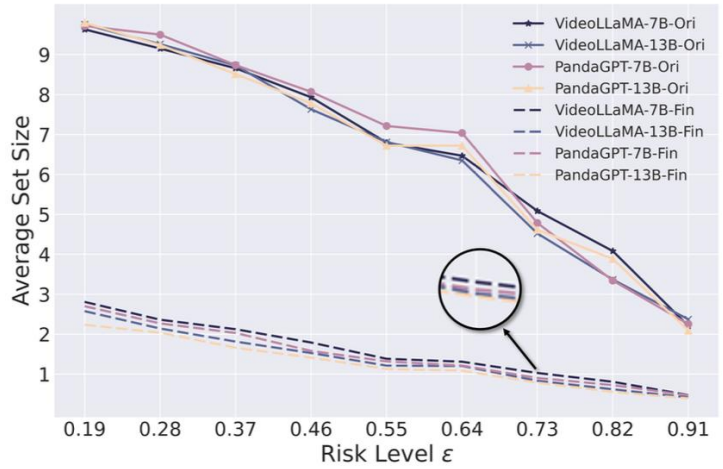
An illustration of our framework in open-domain VideoQA tasks.

Results of *TRON* in Risk Control and Assessment





Results of the EER metric at various risk levels of a and β.

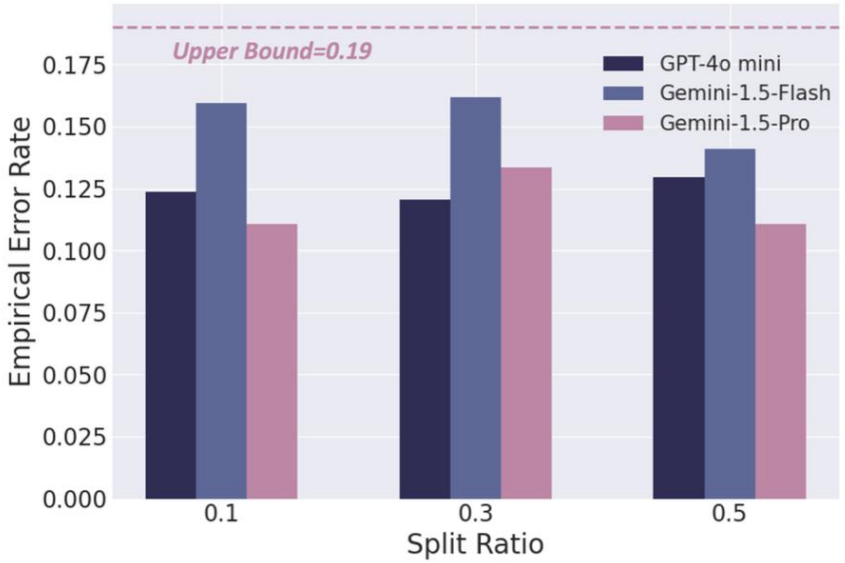


Comparison of APSS before and after Semantic duplication at different risk levels (ε).

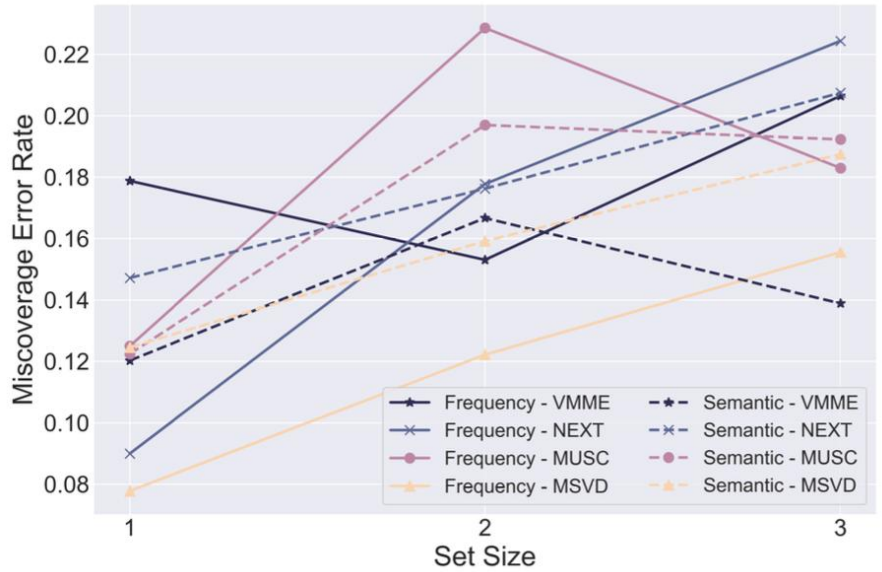
MLLMs	ACC (%) ↑				APSS ↓			
	VMME	NEXT	MUSC	MSVD	VMME	NEXT	MUSC	MSVD
Open-source Models								
VideoLLaMA-7B	38.10	46.06	78.70	78.29	2.81	2.68	1.97	1.63
VideoLLaMA-13B	45.33	47.07	79.18	78.91	2.58	2.37	1.59	1.63
PandaGPT-7B	43.45	50.95	81.85	77.29	2.70	2.10	2.02	1.12
PandaGPT-13B	48.75	58.90	82.73	86.04	2.24	2.20	2.03	2.21
NExT-GPT	42.64	42.59	79.84	84.37	2.43	2.45	2.20	1.70
Closed-source Models								
Gemini-1.5-Flash	72.31	70.80	85.58	87.14	1.22	1.17	1.55	1.08
Gemini-1.5-Pro	73.11	76.29	85.16	86.67	1.15	1.15	1.59	1.10
GPT-4o mini	73.26	82.50	86.27	85.71	1.14	1.08	1.43	1.21

Evaluations of the ACC and APSS metrics utilizing five open-source and three closed- source MLLMs across four open-ended and closed-ended datasets.

Sensitivity Analyses of TRON



EERs on the MUSC dataset utilizing three closed-source MLLMs when varying the split ratio of the calibration and test set.

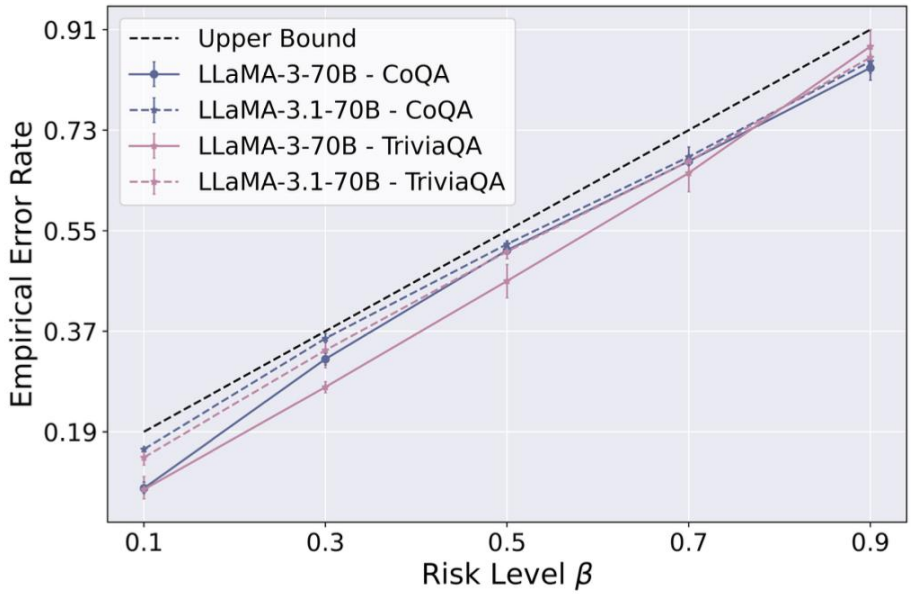


Stratified miscoverage rate at various size of prediction set on four VideoQA datasets utilizing the Gemini-1.5-Pro model.

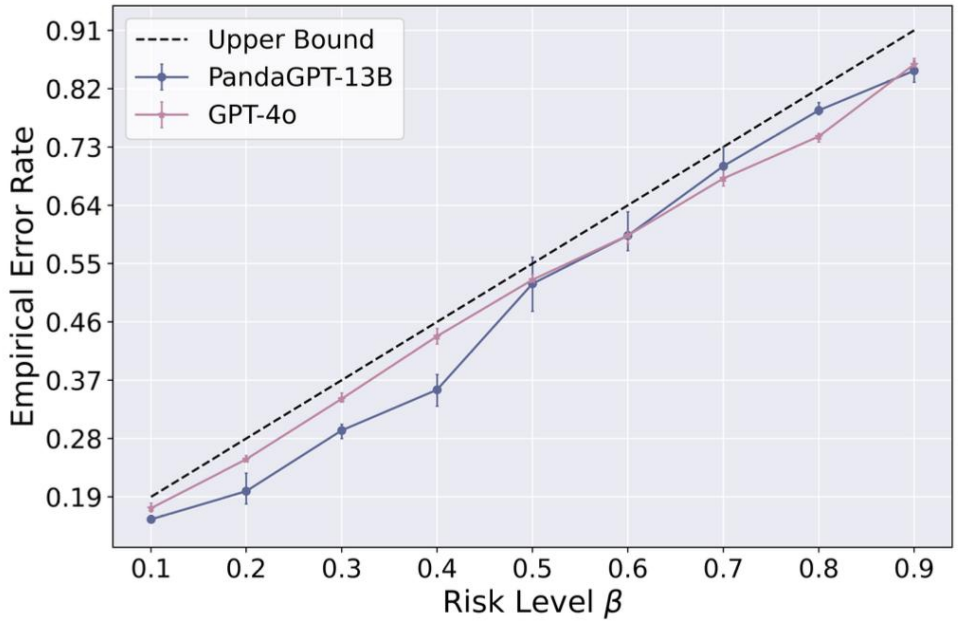
MLLMs	APSS		EER	
	Frequency	Semantic Diversity	Frequency	Semantic Diversity
VideoLLaMA-7B	2.81	2.74	0.1793	0.1766
VideoLLaMA-13B	2.58	2.52	0.1767	0.1744
PandaGPT-7B	2.70	2.42	0.1812	0.1675
PandaGPT-13B	2.24	2.20	0.1783	0.1674

Comparison of frequency and semantic diversity on the VMME datasets employing four open-source MLLMs.

Generalizability of TRON



Results of the EER metric at various risk levels on the CoQA and TriviaQA tasks utilizing LLaMA-3-70B-Instruct and LLaMA-3.1-70B-Instruct.



Results of the EER metric, obtained from 5 trials with randomly allocated calibration and test data at various risk levels on the VQA dataset utilizing PandaGPT-13B and GPT-4o.

Model	*LAC*	*TRON* ($M = 5$)	*TRON* ($M = 10$)	*TRON* ($M = 20$)
LLaMA-3-8B-Instruct	$2.93_{(2)}$	$3.06_{(0)}$	$2.93_{(7)}$	$\mathbf{2.76}_{(3)}$
LLaMA-3.1-8B-Instruct	$2.57_{(8)}$	$2.61_{(3)}$	$2.53_{(6)}$	$\mathbf{2.50}_{(4)}$

The APSS metric (↓) on the MMLU dataset.

In the calibration set, we define the minimum sampling size (conformal score ↑) that ensures $y_i^* \in \left\{y_j^{(i)}\right\}_{j=1}^{M}$ as

$$r_i = r(x_i,\, y_i^*) := \sup\left\{M_i\colon\, \forall M_i^{'} < M_i,\, y_i^* \notin \left\{y_j^{(i)}\right\}_{j=1}^{M_i^{'}}\right\}.$$

We sort the conformal scores so that $r_1 \le \cdots \le r_N$, and calculate the $\frac{\lceil (1-\alpha)(N+1)\rceil}{N}$ quantile:

$$\hat{r} := \inf\left\{r\colon\, \frac{|\{i\colon\, r_i \le r\}|}{N} \ge \frac{\lceil (1-\alpha)(N+1)\rceil}{N}\right\} = r_{\lceil (1-\alpha)(N+1)\rceil}.$$

We proceed by noting the equality of the two events:

$$\left\{y_{N+1}^* \notin \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right\} = \{r_{N+1} > \hat{r}\}.$$

By the exchangeability of $N+1$ data points, we have

$$\mathbb{P}(r_{N+1} \le r_i) = \frac{i}{N+1}.$$

Then, we can guarantee the upper bound of the probability that the sampling set of size $M = \hat{r}$ fails to encompass an acceptable response based on a user-specified risk level of α:

$$\mathbb{P}\left(y_{N+1}^* \notin \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right) = \mathbb{P}(r_{N+1} > \hat{r}) = \mathbb{P}\left(r_{N+1} > r_{\lceil (1-\alpha)(N+1)\rceil}\right)$$

$$= 1 - \mathbb{P}\left(r_{N+1} \le r_{\lceil (1-\alpha)(N+1)\rceil}\right) = 1 - \frac{\lceil (1-\alpha)(N+1)\rceil}{N} \le \alpha$$

If there are no acceptable responses in the candidate set, $\left\{y_{N+1}^* \notin \mathcal{C}\left(\beta,\ \{s_i\}_{i=1}^N,\ x_{N+1},\ \left\{y_j^{(N+1)}\right\}_{j=1}^{M=\hat{r}}\right)\right\}$ is a certain event, and the probability of miscoverage is

$$\mathbb{P}\left(\left\{y_{N+1}^* \notin \mathcal{C}\left(\beta,\ \{s_i\}_{i=1}^N,\ x_{N+1},\ \left\{y_j^{(N+1)}\right\}_{j=1}^{M}\right)\right\} \mid y_{N+1}^* \notin \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right) = \mathbb{P}\left(y_{N+1}^* \notin \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right).$$

Conversely, if acceptable responses exist in the sampling set, we define the non-conformity score (↓).

First, we determine the acceptable response of the i-th calibration data point $y_{ref}^{(i)}$, , which is semantically equivalent to the correct answer y_i^*. Then we formulate the non-conformity score as

$$s_i = s(x_i,\ y_i^*) = 1 - F\left(y_{ref}^{(i)}\right),$$

where $F\left(y_{ref}^{(i)}\right)$ can be any reliable notion of $y_{ref}^{(i)}$ in the sampling set (black-box or white box).

We construct the prediction set following

$$\mathcal{C}\left(\beta,\ \{s_i\}_{i=1}^N,\ x_{N+1},\ \left\{y_j^{(N+1)}\right\}_{j=1}^{M}\right) = \left\{y \in \left\{y_j^{(N+1)}\right\}_{j=1}^{M} :\ 1 - F\left(x_{N+1},\ y,\ \left\{y_j^{(N+1)}\right\}_{j=1}^{M}\right) \le \hat{s}\right\}.$$

Given $y_{N+1}^* \in \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}$, we have

$$\left\{y_{N+1}^* \notin \mathcal{C}\left(\beta,\ \{s_i\}_{i=1}^N,\ x_{N+1},\ \left\{y_j^{(N+1)}\right\}_{j=1}^{M}\right)\right\} = \{s(x_{N+1},\ y_{N+1}^*) > \hat{s}\} = \{s_{N+1} > \hat{s}\}$$

In this case, the probability of miscoverage is

$$\mathbb{P}\left(\left\{y_{N+1}^{*} \notin \mathcal{C}\left(\beta, \{s_i\}_{i=1}^{N}, \ x_{N+1}, \left\{y_j^{(N+1)}\right\}_{j=1}^{M}\right)\right\} \mid y_{N+1}^{*} \in \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right) = \mathbb{P}\left(s_{N+1} > \hat{s} \mid y_{N+1}^{*} \in \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right).$$

We have

$$\mathbb{P}(s_{N+1} > \hat{s}) = 1 - \mathbb{P}(s_{N+1} \leq s_i) = 1 - \frac{i}{N+1}.$$

Finally, we achieve risk control and guarantee the upper bound of the correctness miscoverage rate

$$\begin{aligned}
&\mathbb{P}\left(\left\{y_{N+1}^{*} \notin \mathcal{C}\left(\beta, \{s_i\}_{i=1}^{N}, \ x_{N+1}, \left\{y_j^{(N+1)}\right\}_{j=1}^{M}\right)\right\}\right) \\
&= \mathbb{P}\left(y_{N+1}^{*} \notin \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right) + \mathbb{P}\left(s_{N+1} > \hat{s} \mid y_{N+1}^{*} \in \left\{y_j^{(N+1)}\right\}_{j=1}^{\hat{r}}\right) \\
&= 1 - \frac{\lceil (1-\alpha)(N+1) \rceil}{N} + \left(1 - \frac{\lceil (1-\beta)(N+1) \rceil}{N}\right) \cdot \frac{\lceil (1-\alpha)(N+1) \rceil}{N} \\
&= 1 - \frac{\lceil (1-\alpha)(N+1) \rceil}{N} \cdot \frac{\lceil (1-\beta)(N+1) \rceil}{N} \leq 1 - (1-\alpha)(1-\beta) \\
&\leq \alpha + \beta - \alpha\beta.
\end{aligned}$$

If the sampling set of each test sample contains acceptable responses, $\alpha \to 0$, we have

$$\mathbb{P}\left(\left\{y_{N+1}^{*} \in \mathcal{C}\left(\beta, \{s_i\}_{i=1}^{N}, \ x_{N+1}, \left\{y_j^{(N+1)}\right\}_{j=1}^{M}\right)\right\}\right) \geq 1 - \beta.$$