ICLR
International Conference On Learning Representations

UC SANTA CRUZ UC SANTA BARBARA

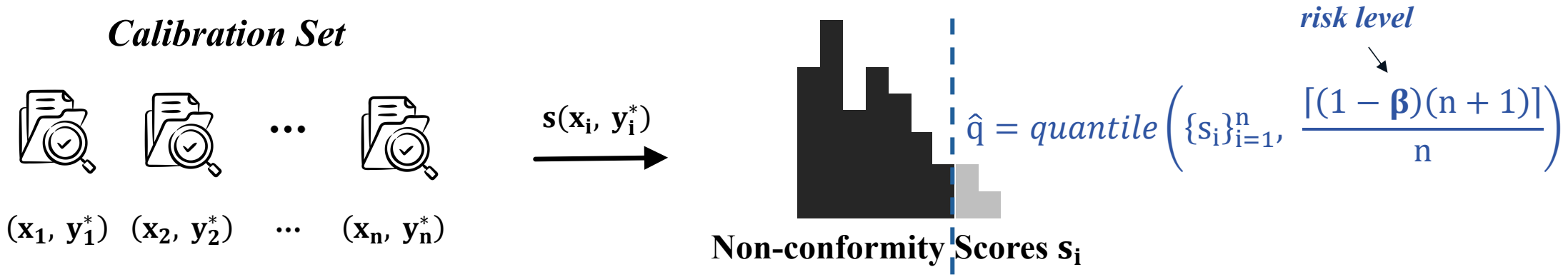
SAFER: Risk-Constrained Sample-then-Filter in Large Language Models

Qingni Wang[1], Yue Fan[1], Xin Eric Wang[1,2]
University of California, Santa Cruz[1] University of California, Santa Barbara[2]

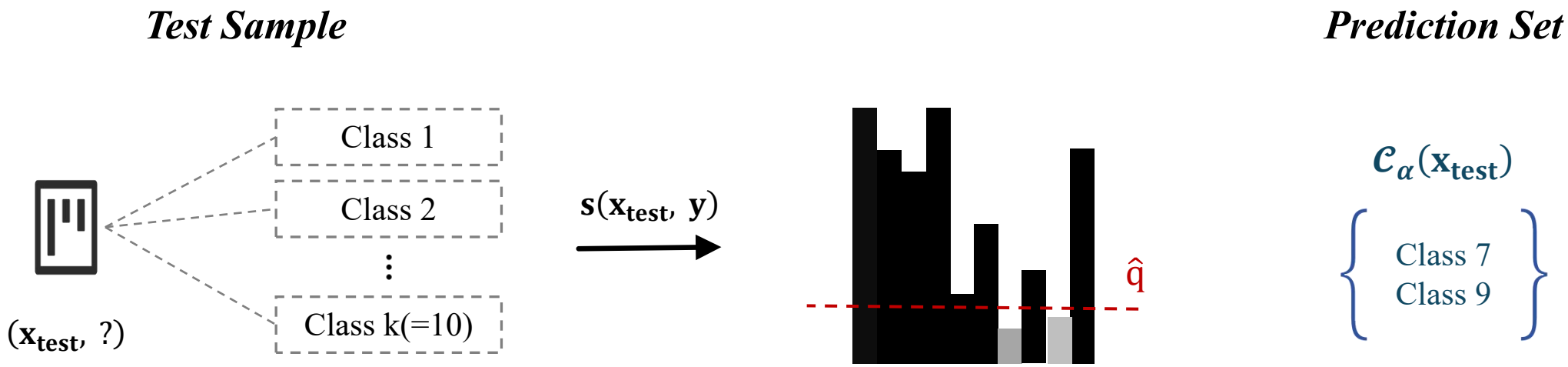
Background of SCP

Split conformal prediction (SCP) offers statistically valid guarantees of **ground-truth coverage** in **classification** tasks.

*SCP measures the **disagreement/residual** between the input x_i and ground-truth label y_i^* on a held-out calibration set, and then calculate the quantile $\hat{q}$ based on a user-specified risk level β.*



The quantile $\hat{q}$ is then fixed to create prediction sets at test time.

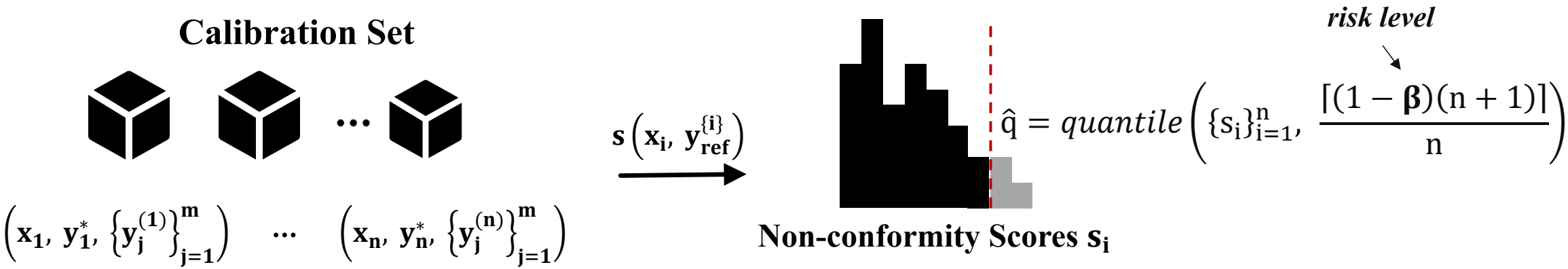


Given a test sample, SCP calculates the ***non-conformity scores*** *between the test sample* $\boldsymbol{x_{test}}$ *and each class and construct the prediction set based on* $\hat{q}$.

On the test set, the average error rate of these prediction sets failing to cover true classes does not exceed $\boldsymbol{\beta}$.

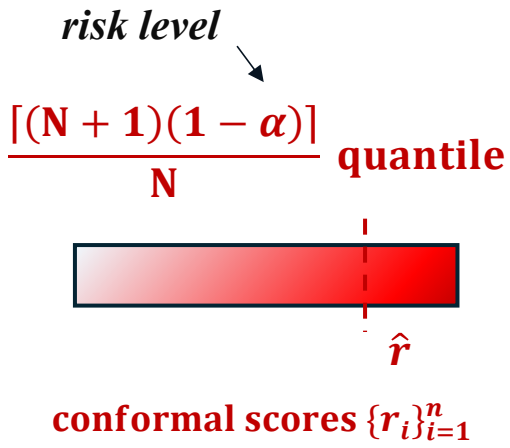
Open-ended tasks do not have a fixed and finite output space, compromising the statistical guarantee!

Limitations of Prior Work



Prior work correlates the non-conformity score with the uncertainty of $\boldsymbol{y}_{ref}^{\{i\}}$ within the sampling set $\left\{\boldsymbol{y}_j^{(i)}\right\}_{j=1}^{m}$, which is semantically equivalent to $\boldsymbol{y}_i^*$ (one minus the frequency score of $\boldsymbol{y}_{ref}^{\{i\}}$), and then employ $\hat{\boldsymbol{q}}$ as the threshold to identify high-quality responses for each test sample.

Limitations of Prior Work



TRON conducts a conformal procedure to manage the average error rate of sampling sets failing to cover admissible responses by developing a ***conformal score*** *that determines the minimum sampling size for each calibration data:*

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

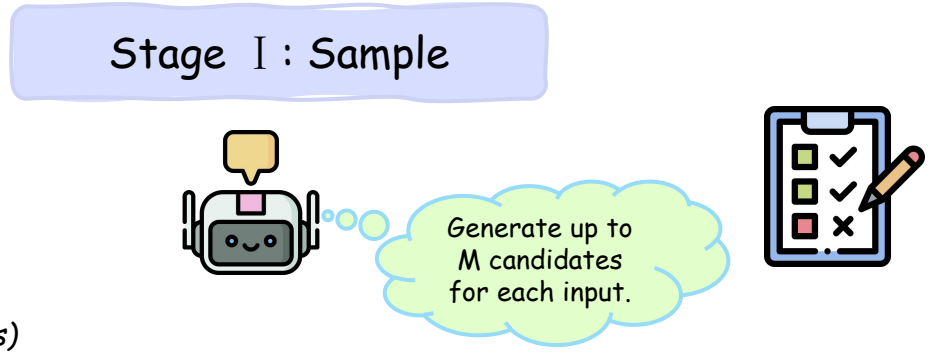
TRON then sets the test-time sampling size to the $\hat{r}$ *based on the other risk level* α.

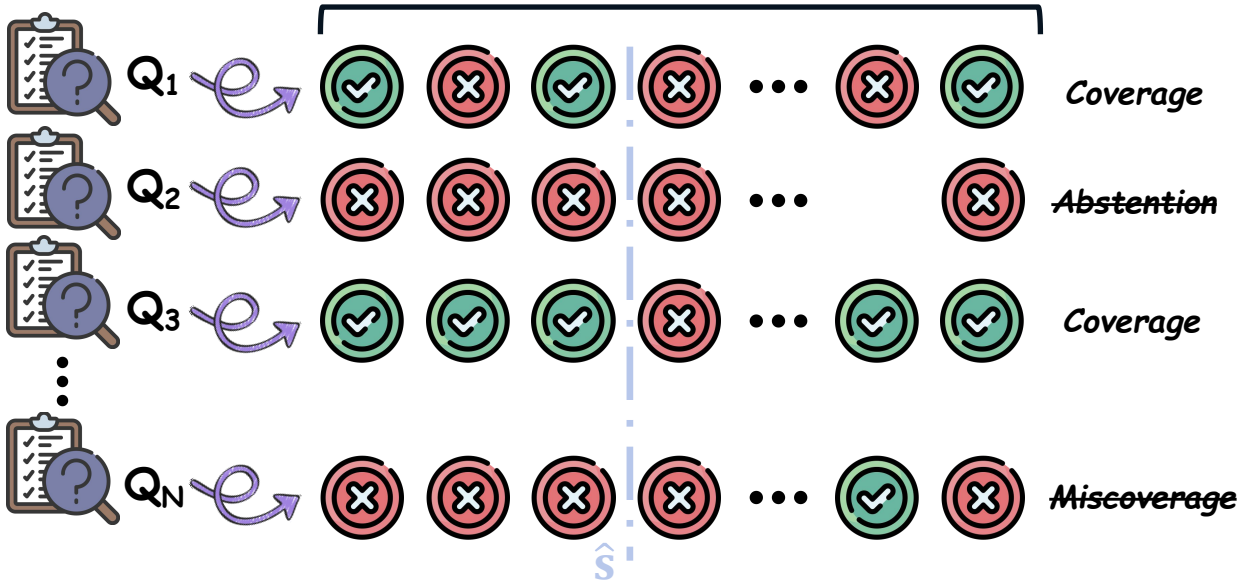
TRON also overlooks that we cannot always ① yield an admissible answer within finite-sampling and ② achieve each user-specified risk level.

Our Method

Goal: (a) determining whether a user-specified risk level is feasible under finite sampling, (b) computing the minimum sampling budget required to meet that risk level.

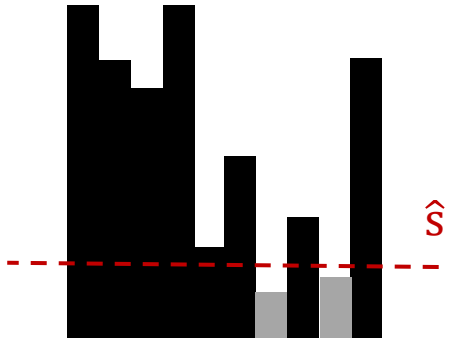
Stage I: How to Sample?





Find minimum budget $\hat{s}$ satisfying user-specified risk α. We abstain this risk level if $\hat{R}^{+}(M) > \alpha$

$$\hat{s} = \inf\{s \in [1, M]: \hat{R}^{+}(s) \leq \alpha\}.$$



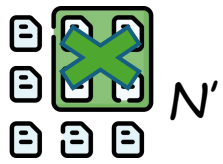
Our Method

Goal: (c) filtering unreliable samples using a statistically calibrated uncertainty threshold.

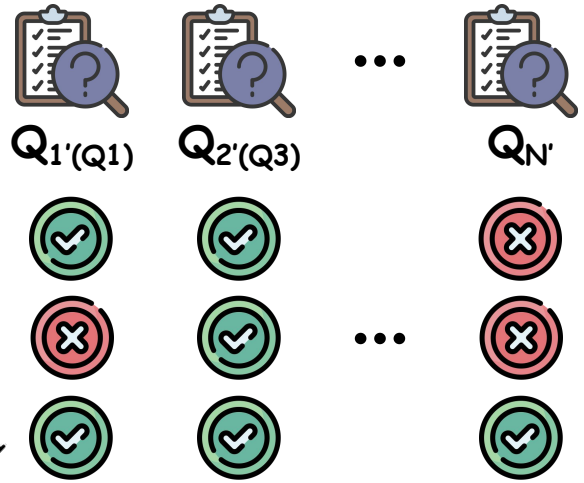
Stage II: How to choose the answers we want?

Stage Ⅱ: Filter

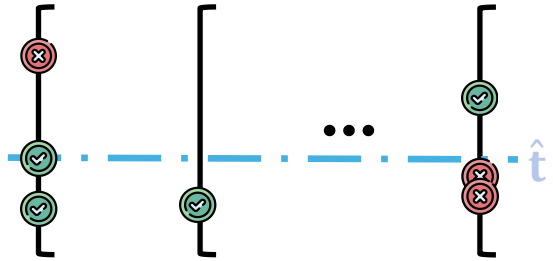
Excluding those never hit within the sampling budget



Calibration Subset
(QA samples with admissible candidates within ŝ generations)



Uncertainty Quantification



Uncertainty Range

And we calibrate uncertainty threshold $\hat{t}$ to control risk β using conformal risk control:

$$\hat{t} = \inf\left\{t: \frac{N' L_{N'}(t) + 1}{N' + 1} \leq \beta\right\}.$$

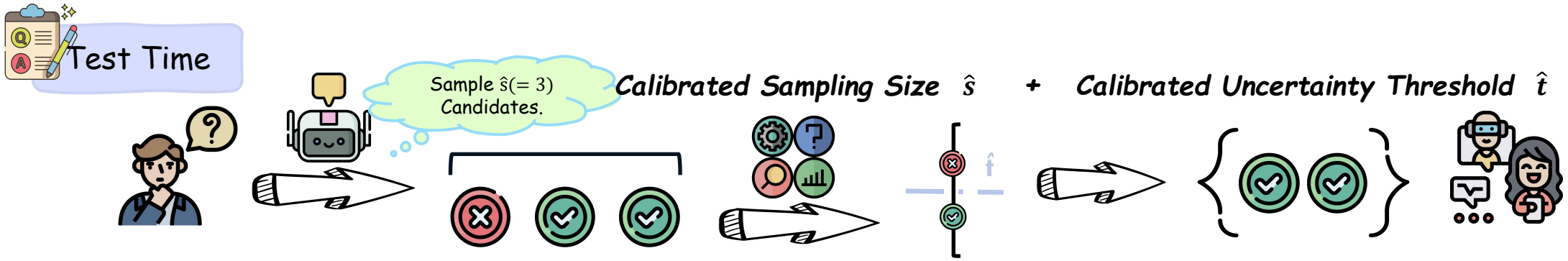
risk level

$$\frac{\lceil (N+1)(1-\beta) \rceil}{N}$$ **quantile**

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

Our Method

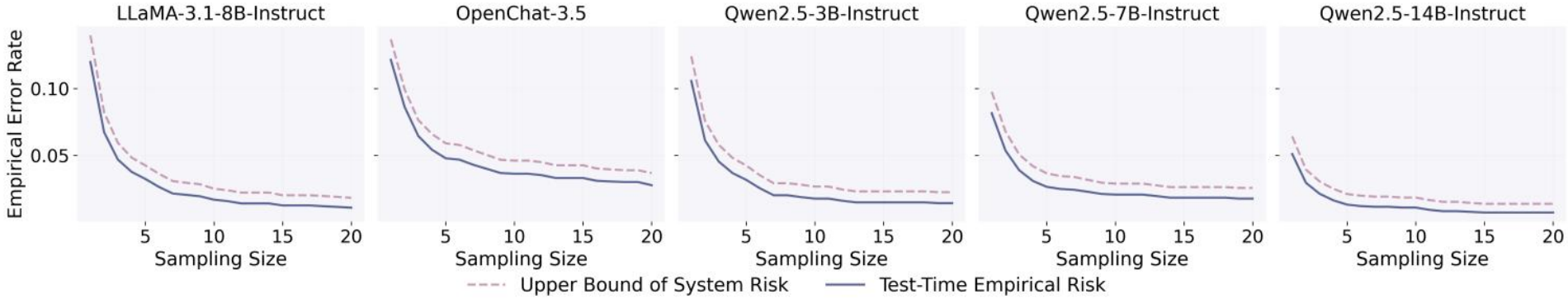
What do we do in the test set? We "learn then test".



Finally, with probability at least $1-\delta$*, the final prediction set* $C_{\hat{t}}(x_{test})$ *satisfies:*

$$\Pr\left(miscoverage \leq \alpha + \beta - \alpha\beta\right).$$

Results

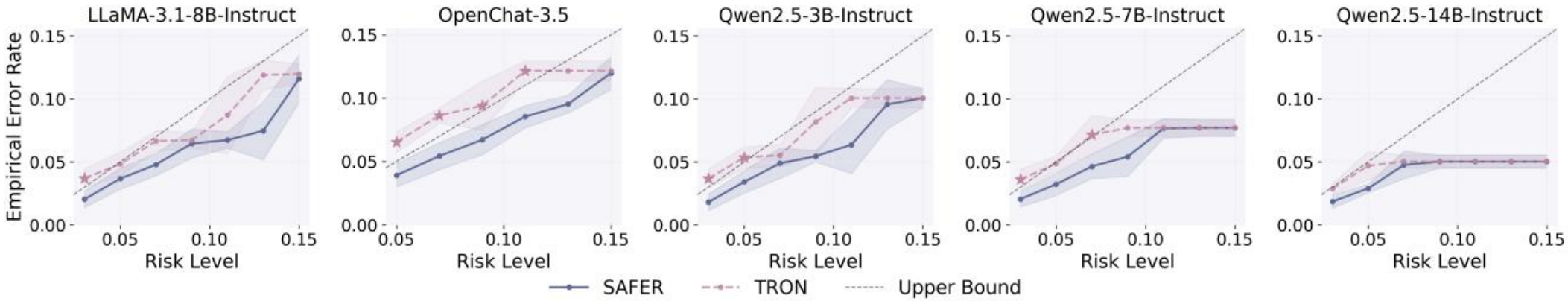


Under each user-specified maximum sampling budget M, SAFER estimates the abstention (i.e., miscoverage) rate on the calibration set, and then calculate the (1-δ) upper confidence bound.

If the risk level fails below the bound, SAFER deem it infeasible, means that we cannot provide statistically valid guarantee at test time, thereby abstaining, highlighting its advantage over point-wise methods ($M = 1$).

Results

SAFER calibrates the sampling budget, enforcing that the upper bound does not exceed the risk level. Compared to TRON, SAFER constrains the test-time error rate at various feasible risk levels.



SAFER also employ the conformal risk control framework to calculate a rigorously calibrated uncertainty threshold to filter out unreliable candidates within the prediction set, while maintaining the statistical guarantee.

β	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**
Upper Bound ($\alpha+\beta-\alpha\beta$)	0.0975	0.1450	0.1925	0.2400	0.2875	0.3350	0.3825	0.4300
LLaMA-3.1-8B-Instruct	0.0844±0.0137	0.1322±0.0160	0.1805±0.0189	0.2265±0.0198	0.2753±0.0215	0.3225±0.0211	0.3698±0.0198	0.4175±0.0195
OpenChat-3.5	0.0890±0.0128	0.1375±0.0155	0.1851±0.0175	0.2318±0.0185	0.2800±0.0211	0.3274±0.0232	0.3752±0.0221	0.4225±0.0214
Qwen2.5-3B-Instruct	0.0822±0.0145	0.1295±0.0189	0.1757±0.0193	0.2223±0.0204	0.2709±0.0212	0.3186±0.0231	0.3683±0.0241	0.4125±0.0236
Qwen2.5-7B-Instruct	0.0802±0.0136	0.1264±0.0158	0.1738±0.0172	0.2207±0.0194	0.2693±0.0209	0.3184±0.0211	0.3663±0.0229	0.4140±0.0241
Qwen2.5-14B-Instruct	0.0771±0.0104	0.1253±0.0146	0.1746±0.0165	0.2223±0.0184	0.2692±0.0212	0.3161±0.0217	0.3612±0.0217	0.4100±0.0221

Test-time EER results in the filtering stage on the TriviaQA (α = 0.05) at various risk levels (β).