MiRD: Reliable Set-Valued Prediction for Open-Ended Question Answering via Miscoverage Risk Decomposition

Anonymous ACL submission

Abstract

Reliable set-valued prediction provides a principled way to mitigate hallucinations in open-ended question answering (QA), yet existing conformal approaches typically rely on a fragile premise: finite sampling must already produce at least one admissible candidate, or calibration examples violating this condition are discarded. In this paper, we introduce **MiRD**, a two-stage framework that decomposes overall miscoverage into *sampling failure* and *conditional selection failure*. In Stage I, MiRD establishes an expectation-level marginal upper bound on the probability that finite sampling produces no admissible answer under a fixed budget. In Stage II, conditioned on sampling success, MiRD calibrates a conformal selection threshold using admission-correlated nonconformity scores defined over the full calibration set, thereby preserving calibration-set integrity. Across three open-ended QA datasets and eight models, MiRD controls sampling risk, conditional selection risk, and overall miscoverage, while yielding tighter first-stage bounds than PAC-style alternatives and more adaptive prediction sets than successful-only calibration.

1 Introduction

Large language models (LLMs) have become a central component of open-ended question answering (QA) systems (Sui et al., 2025; Duan et al., 2025), yet their generated answers remain unreliable (Wang et al., 2025d; Huang et al., 2024): even strong LLMs can produce hallucinated or semantically misaligned responses. A natural route toward reliability is split conformal prediction (SCP), which replaces a single answer with a prediction set that is ensured to contain at least one admissible candidate with high probability (Campos et al., 2024; Zhou et al., 2025). However, such a guarantee implicitly presumes that the candidate pool to be filtered already contains an admissible answer. This premise is nontrivial in open-ended QA, where the output space is unbounded, correctness must be defined semantically, and a finite number of stochastic samples may fail to produce any admissible answer at all (Wang et al., 2025a, 2026a).

This observation suggests that reliability in open-ended QA should not be viewed merely as a post-hoc filtering problem over a pre-existing candidate set. Even when an admissible answer exists in principle, finite sampling may fail to generate it (Wang et al., 2025b); even when such an answer is sampled, subsequent conformalized selection may remove it. Existing conformal methods typically focus on only the latter source of error: they either assume that every sampled candidate set contains at least one admissible answer (Wang et al., 2024b, 2025a), or discard calibration data points that violate this condition (Kaur et al., 2024), thereby restricting calibration to a success-conditioned subset of the data distribution. More recent sample-then-filter approaches relax this assumption using confidence-interval-based procedures (Wang et al., 2026a), but establish PAC-style guarantees with an additional confidence parameter rather than a direct marginal conformal guarantee. These limitations motivate a framework that explicitly separates *finite-sampling uncertainty* from *conditional post-selection uncertainty*, while preserving the integrity of the full calibration set (Wang et al., 2025e).

In this work, we propose **MiRD**, a **Mi**scoverage **R**isk **D**ecomposition framework for reliable set-valued prediction in open-ended QA. The key idea is to decompose overall miscoverage into two statistically distinct sources: (i) a *marginal sampling-failure risk*, where finite stochastic sampling fails to produce any admissible answer, and (ii) a *conditional selection-failure risk*, where conformalized filtering removes all admissible answers despite their presence in the sampled candidate set. This decomposition naturally leads to a two-stage procedure. In Stage I, we derive an expectation-level marginal upper bound on the finite-sampling fail-

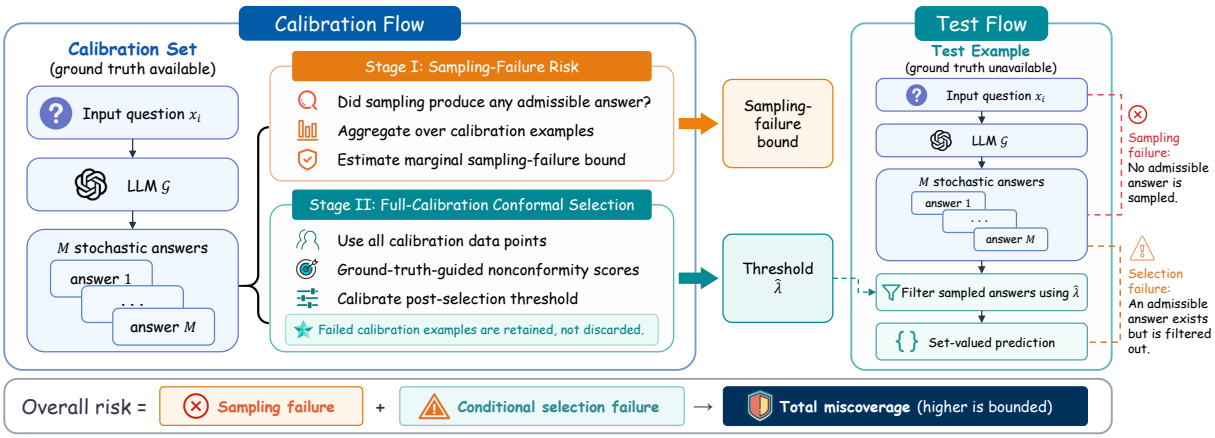


Figure 1: Overview of two-stage miscoverage risk decomposition.

ure risk under a fixed sampling budget. In Stage II, conditional on sampling success, we calibrate a conformal selection threshold using admission-correlated nonconformity scores defined over the full calibration set. Crucially, MiRD does not discard calibration data whose sampled candidates contain no admissible answer. Instead, it leverages the availability of ground-truth answers during calibration to construct score information for all calibration examples, thereby preserving calibration-set integrity while enabling valid threshold calibration. An overview of MiRD is shown in Figure 1.

This decomposition offers both conceptual and practical benefits. Conceptually, it separates open-ended miscoverage into finite-sampling failure and post-selection failure. Practically, it controls sampling failure explicitly and avoids the distributional distortion introduced by successful-only calibration. Although the selection stage relies on a score-level conditional compatibility assumption, we further show that, under our uncertainty construction, full-calibration thresholding can be formulated as a monotone enlargement of successful-only calibration: retaining sampling-failed calibration data yields a weakly more conservative threshold and thus a superset predictor on sampling-successful test data. This explains why MiRD preserves conformal validity while using the full calibration set.

We evaluate MiRD on three QA datasets with eight LLMs from four widely utilized model families. MiRD consistently controls finite-sampling risk, conditional selection risk, and overall miscoverage risk. Compared with PAC-style confidence bounds (Bates et al., 2021), MiRD provides substantially tighter first-stage bounds while maintaining empirical validity. Compared with successful-only calibration (Wang et al., 2024b, 2025a; Kaur et al., 2024), MiRD yields prediction sets that better adapt to example difficulty. Overall, these results show that miscoverage risk decomposition benefits both statistical validity and the practical utility of set-valued prediction in open-ended generation.

In summary, our contributions are threefold:

- We reframe reliable open-ended QA as a miscoverage risk decomposition problem, separating overall failure into marginal sampling failure and conditional selection failure.
- We introduce MiRD, a two-stage calibration framework that combines an expectation-level marginal upper bound for finite-sampling risk with full-calibration conformalized selection for conditional post-selection risk control.
- We validate MiRD across multiple LLM families and QA benchmarks, achieving tighter sampling-risk bounds, more adaptive prediction sets, and competitive end-to-end miscoverage compared with prior alternatives.

2 Related Work

Split Conformal Prediction. SCP converts arbitrary nonconformity scores into prediction sets with finite-sample marginal coverage guarantees under exchangeability (Papadopoulos, 2008; Angelopoulos and Bates, 2023; Tan et al., 2025). Given a held-out calibration set, SCP calibrates a threshold from empirical nonconformity scores and applies it to future test examples according to a user-specified risk level. Recent work has adapted SCP to LLMs, but largely focuses on closed-ended scenarios (Kumar et al., 2023; Cresswell et al., 2024; Ye et al., 2024) or output modification (Cherian et al., 2024). Reliable set-valued prediction for open-ended QA remains less settled, because admissible answers must be obtained from a finite sampled candidate pool rather than selected from a fixed label space.

SCP in Open-Ended QA. For open-ended generation, Kaur et al. (2024) defines nonconformity as the uncertainty of admissible semantics within sampled answers (Kuhn et al., 2023). ConU targets API-only settings and constructs self-consistency-based nonconformity, but assumes that every sampled set contains admissible answers (Wang et al., 2024b). TRON further develops a two-stage risk control framework that combines sampling-size calibration with conformal selection, while relying on the same sampling-success premise for both calibration and test data (Wang et al., 2025a). Quach et al. (2024) and SAFER (Wang et al., 2026a) relax this premise through confidence-interval-based sample-then-filter procedures, but offer PAC-style guarantees with an additional confidence parameter δ. In contrast, MiRD explicitly decomposes overall miscoverage into marginal finite-sampling failure and conditional post-selection failure, and performs full-calibration conformalized filtering without discarding sampling-failed calibration points.

See additional related work in Appendix B.

3 Methodology

3.1 Notations and Problem Formulation

Let $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ be a pretrained LLM that maps an input prompt $x \in \mathcal{X}$ to a textual answer $\hat{y} \in \mathcal{Y}$. Following standard split conformal protocols (Angelopoulos et al., 2024), we hold out a calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i^*)\}_{i=1}^N$, where $x_i \in \mathcal{X}$ is the i-th prompt and $y_i^* \in \mathcal{Y}$ is the corresponding ground truth. We consider one test example (x_{N+1}, y_{N+1}^*), where y_{N+1}^* is unavailable to the user. We assume that the calibration examples together with the test point are exchangeable. For each input x, we draw M stochastic answers from $\mathcal{G}$ to form a finite candidate set $\mathcal{G}_M(x) = \{\hat{y}_m\}_{m=1}^M$. For calibration and test examples, we write $\mathcal{G}_M(x_i) = \{\hat{y}_m^{(i)}\}_{m=1}^M$ for all $i \in \{1, \cdots, N+1\}$. To evaluate answer correctness, let $A : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ be a binary admission function, where $A(y^*, \hat{y}) = 1$ indicates that $\hat{y}$ is admissible with respect to the ground-truth answer y^*, and $A(y^*, \hat{y}) = 0$ otherwise. Note that the admission rule can be instantiated by any task-specific semantic equivalence criterion (e.g., bi-entailment), and our framework is agnostic to its specific form.

A central challenge in open-ended generation is that finite sampling does not ensure the appearance of an admissible answer. To capture this irreducible source of error, we define the sampling failure indicator for the test example as

$$Z_{N+1} = \mathbf{1}\big(\forall \hat{y} \in \mathcal{G}_M(x_{N+1}), A(y_{N+1}^*, \hat{y}) = 0\big)\,, \tag{1}$$

which equals 1 if no admissible answer is observed among the M candidates for the future test input x_{N+1}, and 0 otherwise. We denote the corresponding sampling-failure probability by

$$p_{\text{fail}} := \Pr(Z_{N+1} = 1). \tag{2}$$

Our first objective is to characterize this quantity through a marginal upper bound in expectation.

Conditioned on the event $\{Z_{N+1} = 0\}$, i.e., the sampling set containing at least one admissible answer, we then perform conformalized selection. Let $U_{\mathcal{G}}(\hat{y} \mid x) \in \mathbb{R}$ be an uncertainty score assigned by model $\mathcal{G}$ to the candidate $\hat{y}$ under input x. Given a threshold λ, we define the filtered prediction set as

$$\mathcal{C}_\lambda(x_{N+1}) = \{\hat{y} \in \mathcal{G}_M(x_{N+1}) : U_{\mathcal{G}}(\hat{y} \,|\, x_{N+1}) \le \lambda\}\,. \tag{3}$$

To formalize failure after filtering, we then define the miscoverage indicator as

$$R_{N+1}(\lambda) = \mathbf{1}\big(\forall \hat{y} \in \mathcal{C}_\lambda(x_{N+1}), A(y_{N+1}^*, \hat{y}) = 0\big)\,. \tag{4}$$

Our second objective is to calibrate a threshold $\hat{\lambda}$ such that

$$\Pr\Big(R_{N+1}(\hat{\lambda}) = 1 \mid Z_{N+1} = 0\Big) \le \alpha, \tag{5}$$

guided by a user-specified risk level $\alpha \in (0, 1)$.

Given that event $\{Z_{N+1} = 0\}$ is unobservable at test time, constraining the conditional risk in Eq. (5) is insufficient to guarantee the reliability of the final prediction sets. **Our overall goal** is therefore to construct coverage-valid set-valued predictions for future test inputs by explicitly decomposing the overall miscoverage risk into two sources: (i) the marginal risk that finite sampling produces no admissible answer at all, and (ii) the conditional risk that conformalized selection removes all admissible answers when such answers do exist.

To this end, we decompose the overall risk:

$$\begin{aligned}\Pr(R_{N+1}(\lambda) = 1) = \Pr(Z_{N+1} = 1) + \\ \Pr(R_{N+1}(\lambda) = 1 \,|\, Z_{N+1} = 0)\Pr(Z_{N+1} = 0).\end{aligned} \tag{6}$$

Eq. (6) motivates our two-stage framework: we first upper-bound the marginal sampling risk induced by finite generation, and then calibrate $\hat{\lambda}$ to control the conditional selection risk given $Z_{N+1} = 0$.

3.2 Bounding Sampling-Failure Probability

The conditional guarantee in Eq. (5) only applies to test instances for which finite sampling has already

produced at least one admissible answer. We therefore first quantify the marginal probability that a fixed sampling budget M fails to generate any admissible answer at all. Since our goal is to control the average failure rate over future test data, rather than to identify which particular instances will fail, we target a marginal guarantee in expectation.

To make the dependence on the sampling budget explicit, for each data point (x_i, y_i^*), we define

$$Z_i(M) = \mathbf{1}(\forall \hat{y} \in \mathcal{G}_M(x_i),\ A(y_i^*, \hat{y}) = 0)\,. \quad (7)$$

Thus, $\{Z_i(M) = 1\}$ indicates that no admissible answer is observed within M stochastic sampling. We assume that the candidate sets are generated sequentially, so that $\mathcal{G}_{M_1}(x) \subseteq \mathcal{G}_{M_2}(x)$ whenever $M_1 \leq M_2$. Under the construction, $Z_i(M)$ is non-increasing in M and bounded in $[0, 1]$.

Given a fixed sampling budget M, we define the empirical sampling-failure rate on average over the calibration set as

$$\widehat{R}_N(M) = \frac{1}{N}\sum_{i=1}^{N} Z_i(M). \quad (8)$$

Since $Z_i(M) \in \{0, 1\}$, the sampling-failure probability for future test points satisfies

$$p_{\text{fail}}(M) = \Pr\big(Z_{N+1}(M) = 1\big) = \mathbb{E}[Z_{N+1}(M)], \quad (9)$$

which means, for this binary loss, the failure probability and the expected loss coincide (Angelopoulos and Bates, 2023; Angelopoulos et al., 2024).

The following proposition connects the test-time failure probability to the empirical sampling-failure rate measured on the calibration set.

Proposition 1 (Marginal upper bound for a fixed sampling budget)**.** *Assume that the N calibration data points and one future test data are exchangeable. Then, for any fixed sampling budget M,*

$$\begin{aligned} p_{\text{fail}}(M) &= \Pr\big(Z_{N+1}(M) = 1\big) \\ &\leq \mathbb{E}\left[\frac{N}{N+1}\widehat{R}_N(M) + \frac{1}{N+1}\right]. \end{aligned} \quad (10)$$

See the proof of Proposition 1 in Appendix A.

Proposition 1 establishes a marginal upper bound for the first-stage sampling risk under the fixed budget M. Importantly, the guarantee is in expectation over both the calibration instances and the future test data point. Therefore, unlike prior PAC-style approaches (Wang et al., 2026a), the quantity

$$\widetilde{\beta}(M) := \frac{N}{N+1}\widehat{R}_N(M) + \frac{1}{N+1} \quad (11)$$

can be interpreted as a calibration-dependent proxy whose expectation upper-bounds test-time marginal failure probability, rather than as a high-probability (i.e., $\geq 1-\delta$) upper confidence bound on a fixed but unknown parameter (Clopper and Pearson, 1934).

At test time, we cannot determine which particular data satisfy $Z_{N+1}(M) = 0$. However, Proposition 1 guarantees that, *on average over future test points*, the fraction of instances for which finite-sampling fails to yield an admissible answer is controlled by the exchangeability-corrected empirical risk in Eq. (11). In the subsequent section, conditioned on the complementary event $Z_{N+1} = 0$, we calibrate the conformalized selection threshold to control the second-stage conditional risk.

3.3 Conformalized Selection

Even when finite generation fails to produce an admissible answer, the ground-truth answer remains available for every calibration data point. Given a calibration input, we therefore estimate the model uncertainty assigned to the ground truth to define the nonconformity score (NS). For each calibration example (x_i, y_i^*), we formulate

$$u_i := U_{\mathcal{G}}(y_i^* \mid x_i), \qquad i = 1, \ldots, N. \quad (12)$$

In this way, we establish the admission-correlated nonconformity for *all* calibration data. This avoids discarding calibration instances whose candidate sets fail to cover an admissible answer (Wang et al., 2024b, 2025a; Kaur et al., 2024), thereby preserving the integrity (distribution) of calibration data.

Let $\{u_{(1)} \leq \cdots \leq u_{(N)}\}$ represent the ordered calibration NSs, and define

$$k_\alpha := \lceil (N+1)(1-\alpha) \rceil\,. \quad (13)$$

We calibrate the selection threshold as the conformal quantile

$$\hat{\lambda} := u_{(k_\alpha)}, \quad (14)$$

with the standard convention that $u_{(N+1)} := +\infty$.

Now consider a test data point x_{N+1} satisfying $Z_{N+1} = 0$, i.e., its candidate set contains at least one admissible answer. Among all admissible candidates, we define the reference admissible answer:

$$\hat{y}_{N+1}^{(\text{ref})} \in \underset{\hat{y} \in \mathcal{G}_M(x_{N+1}):\, A(y_{N+1}^*, \hat{y}) = 1}{\arg\min} U_{\mathcal{G}}(\hat{y} \mid x_{N+1}), \quad (15)$$

and let its associated admission-correlated NS be

$$u_{N+1} := U_{\mathcal{G}}(\hat{y}_{N+1}^{(\text{ref})} \mid x_{N+1}). \quad (16)$$

By construction, on event $\{Z_{N+1} = 0\}$, we have

$$\{\exists \hat{y} \in \mathcal{C}_{\hat{\lambda}}(x_{N+1}), A(y^*_{N+1}, \hat{y}) = 1\} = \{u_{N+1} \leq \hat{\lambda}\}. \tag{17}$$

We further impose the exchangeability requirement for this stage: conditional on $Z_{N+1} = 0$, the augmented score sequence $\{u_1, \ldots, u_N, u_{N+1}\}$ is exchangeable. That is, the admission-correlated NS of the test admissible answer is distributed symmetrically with the calibration-side ground-truth NSs. We discuss this assumption in Appendix A.3. Under this formulation, the standard split conformal quantile argument yields

$$\begin{aligned} &\Pr\left(u_{N+1} \leq \hat{\lambda} \mid Z_{N+1} = 0\right) \\ &= \Pr\left(u_{N+1} \leq u_{(k_\alpha)} \mid Z_{N+1} = 0\right) \\ &\geq \frac{k_\alpha}{N+1} \geq 1 - \alpha. \end{aligned} \tag{18}$$

Combining Eqs. (17) and (18), we obtain

$$\begin{aligned} &\Pr\left(\exists \hat{y} \in \mathcal{C}_{\hat{\lambda}}(x_{N+1}), A(y^*_{N+1}, \hat{y}) = 1 \middle| Z_{N+1} = 0\right) \\ &\qquad \geq 1 - \alpha, \end{aligned} \tag{19}$$

and equivalently,

$$\Pr\left(R_{N+1}(\hat{\lambda}) = 1 \mid Z_{N+1} = 0\right) \leq \alpha. \tag{20}$$

Therefore, Eq. (5) holds with the threshold $\hat{\lambda}$.

3.4 Overall Coverage Guarantee

We combine the marginal sampling-failure bound in Proposition 1 with the conditional guarantee in Eq. (20). For a future test example, the decomposition in Eq. (6) gives

$$\begin{aligned} \Pr\left(R_{N+1}(\hat{\lambda}) = 1\right) = \Pr(Z_{N+1}(M) = 1) + \\ \Pr\left(R_{N+1}(\hat{\lambda}) = 1 \mid Z_{N+1}(M) = 0\right) \cdot \\ \Pr(Z_{N+1}(M) = 0). \end{aligned} \tag{21}$$

Using Proposition 1 and Eq. (20), we obtain

$$\begin{aligned} \Pr\left(R_{N+1}(\hat{\lambda}) = 1\right) &\leq p_{\text{fail}}(M) + \alpha \\ &\leq \mathbb{E}\left[\widetilde{\beta}(M)\right] + \alpha \end{aligned}. \tag{22}$$

Therefore, MiRD establishes an expectation-level marginal coverage guarantee for open-ended set-valued prediction. Eq. (22) is conservative; Appendix A.4 gives a tighter population-level refinement by exploiting the complementarity between sampling failure and conditional selection failure.

4 Experiments

4.1 Experimental Setup

Datasets. We use the closed-book TriviaQA (Joshi et al., 2017) and open-book CoQA (Reddy et al., 2019) datasets, covering both knowledge-intensive and context-supported generation settings to test the robustness of MiRD. We also consider more challenging Natural Questions (NQ) (Kwiatkowski et al., 2019). See details in Appendix C.

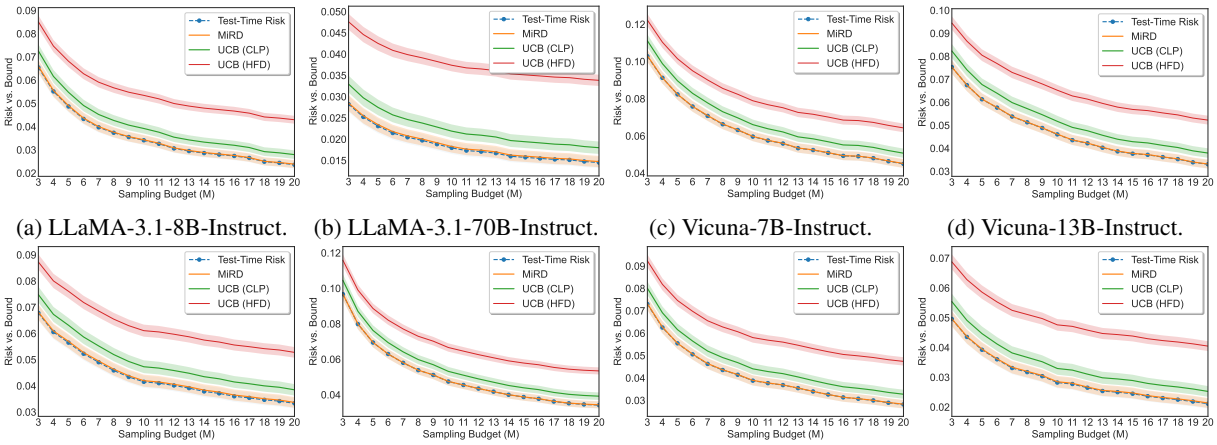
Models. We use eight open-source LLMs, including four popular families: LLaMA (Touvron et al., 2023) (LLaMA-3.1-8B-Instruct and LLaMA-3.1-70B-Instruct), Qwen (Bai et al., 2023) (Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct), Vicuna (Zheng et al., 2023) (Vicuna-7B-Instruct and Vicuna-13B-Instruct), and OpenChat (Wang et al., 2024a) (OpenChat-3.5).

Baselines. In the sampling stage, previous studies typically construct a $(1-\delta)$-level upper confidence bound (UCB) on the sampling-size-conditioned system failure rate (Wang et al., 2026a). We compare our marginal upper bound in Proposition 1 with both the Clopper–Pearson-style (Clopper and Pearson, 1934) and Hoeffding-style (Hoeffding, 1963) UCBs, denoted as UCB-CLP and UCB-HFD, respectively. In the conformal selection stage, we compare against ConU (Wang et al., 2024b), which discards calibration examples whose sampling sets fail to contain admissible answers and calibrates the threshold only on the remaining subset. This comparison directly tests the benefit of preserving full calibration-set integrity in MiRD.

Correctness Criteria. We use sentence similarity by default. To examine admission-sensitivity, we also use bi-entailment. See details in Appendix C.

Uncertainty Measure. Following previous studies (Wang et al., 2024b, 2025a, 2026a), we perform semantic clustering over the candidate set (Kuhn et al., 2023) and define the uncertainty score of each candidate as one minus its cluster frequency. This yields a normalized uncertainty measure $U(\cdot) \in [0, 1]$, where larger estimates represent weaker self-consistency and therefore lower reliability.

Evaluation Criteria. Following prior studies (Angelopoulos and Bates, 2023; Quach et al., 2024; Wang et al., 2025a, 2026a), we randomly split the calibration and test set 500 times, and calculate the mean of test-time failure (miscoverage) proportion as the Monte Carlo approximation of failure (miscoverage) rate. As long as the means do not exceed the corresponding upper bounds, statistical validity



(a) LLaMA-3.1-8B-Instruct. (b) LLaMA-3.1-70B-Instruct. (c) Vicuna-7B-Instruct. (d) Vicuna-13B-Instruct.

(e) OpenChat-3.5 (7B). (f) Qwen2.5-3B-Instruct. (g) Qwen2.5-7B-Instruct. (h) Qwen2.5-14B-Instruct.

Figure 2: Sampling risk vs. three upper bounds at various sampling budgets (M) on TriviaQA with eight LLMs.

Table 1: Sampling risk vs. bound on TriviaQA with the LLaMA-3.1-8B-Instruct model, using sentence similarity for correctness evaluation across various thresholds.

Threshold / M		5	10	15	20
0.5	Risk	0.0107	0.0067	0.0058	0.0044
	Bound	**0.0110**	**0.0069**	**0.0061**	**0.0047**
0.55	Risk	0.0249	0.0165	0.0138	0.0112
	Bound	**0.0252**	**0.0167**	**0.0141**	**0.0114**
0.6	Risk	0.0486	0.0340	0.0279	0.0235
	Bound	**0.0491**	**0.0344**	**0.0282**	**0.0239**
0.65	Risk	0.0876	0.0632	0.0533	0.0468
	Bound	**0.0880**	**0.0635**	**0.0536**	**0.0471**
0.7	Risk	0.1349	0.1006	0.0855	0.0757
	Bound	**0.1355**	**0.1010**	**0.0858**	**0.0760**

Table 2: Sampling risk vs. bound on TriviaQA with the LLaMA-3.1-Instruct family of two model sizes, using bi-entailment for correctness evaluation.

Model Size / M		5	10	15	20
8B	Risk	0.1779	0.1381	0.1196	0.1073
	Bound	**0.1782**	**0.1383**	**0.1198**	**0.1074**
70B	Risk	0.0693	0.0557	0.0475	0.0450
	Bound	**0.0696**	**0.0560**	**0.0479**	**0.0454**

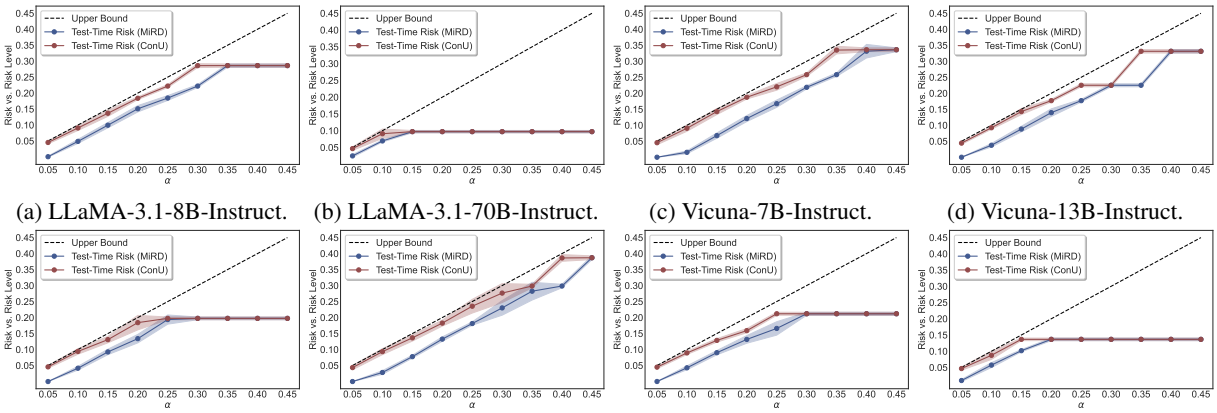
holds. See Angelopoulos et al. (2024) for details.

Evaluation Protocol. Our theory targets *marginal* validity, so all empirical risks are evaluated by repeated random calibration-test splits (Angelopoulos and Bates, 2023; Quach et al., 2024; Wang et al., 2025a, 2026a). We randomly split the data into calibration and test sets 500 times and report the mean test-time risk across splits as a Monte Carlo approximation of the corresponding population quantity. We evaluate three risks corresponding to the three layers of our framework. In the sampling stage, we report the *sampling-failure risk*, i.e., the average proportion of test examples for which no admissible answer is obtained within a fixed sampling budget. In the conformalized selection stage, we report the *conditional selection risk*, i.e., the average miscoverage rate on the subset of test data satisfying $Z_{N+1}(M) = 0$. For the final prediction sets, we report the *overall miscoverage risk*, i.e., the average rate at which the final set fails to contain any admissible answer over all test data points. We regard a statistical guarantee as empirically validated when the corresponding Monte Carlo risk estimate does not exceed its theoretical upper bound.

4.2 Statistical Validity of the Marginal Upper Bound for Finite-Sampling Risk Control

We first evaluate the statistical validity of Proposition 1. Since our first-stage guarantee is a *marginal upper bound in expectation*, rather than a high-probability confidence bound on a fixed parameter, we follow the repeated-splitting protocol in Section 4.1 and estimate both the test-time sampling-failure risk and the bound by Monte Carlo averaging over 500 random calibration-test splits. Concretely, for each split and each fixed sampling budget M, we compute the test-time proportion of examples whose candidate sets fail to contain any admissible answer, together with the exchangeability-corrected empirical proxy $\widetilde{\beta}(M)$, and then report their averages across splits as approximations of $p_{\text{fail}}(M)$ and $\mathbb{E}[\widetilde{\beta}(M)]$, respectively.

As illustrated in Figure 2, across all eight LLMs and all sampling budgets considered, the estimated MiRD bound consistently stays above the empirical sampling-failure risk on TriviaQA, confirming the statistical validity of our marginal upper bound in practice. At the same time, the MiRD bound is uniformly below both UCB-CLP and UCB-HFD,



(a) LLaMA-3.1-8B-Instruct. (b) LLaMA-3.1-70B-Instruct. (c) Vicuna-7B-Instruct. (d) Vicuna-13B-Instruct.

(e) OpenChat-3.5 (7B). (f) Qwen2.5-3B-Instruct. (g) Qwen2.5-7B-Instruct. (h) Qwen2.5-14B-Instruct.

Figure 3: Conditional selection risk vs. upper bound at various risk levels on TriviaQA with eight LLMs ($M = 5$).

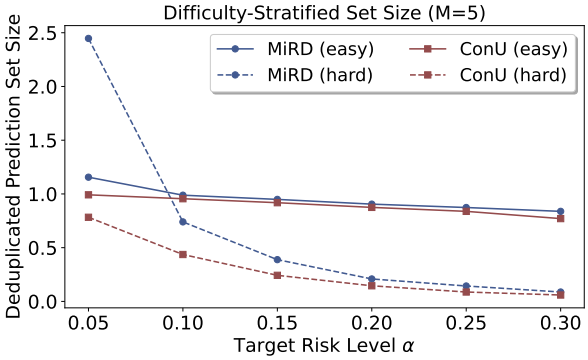


Figure 4: Difficulty-stratified deduplicated prediction set size on TriviaQA with Vicuna-7B-Instruct. Test samples are partitioned into easy and hard groups according to whether the top-1 candidate is admissible.

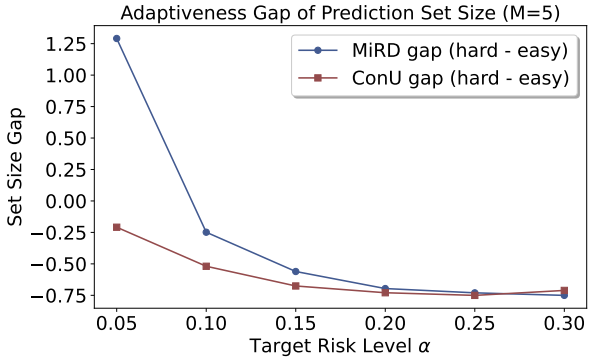


Figure 5: Adaptiveness gap of prediction set size on TriviaQA with Vicuna-7B-Instruct, defined as the difference between the average deduplicated prediction set sizes of hard and easy examples.

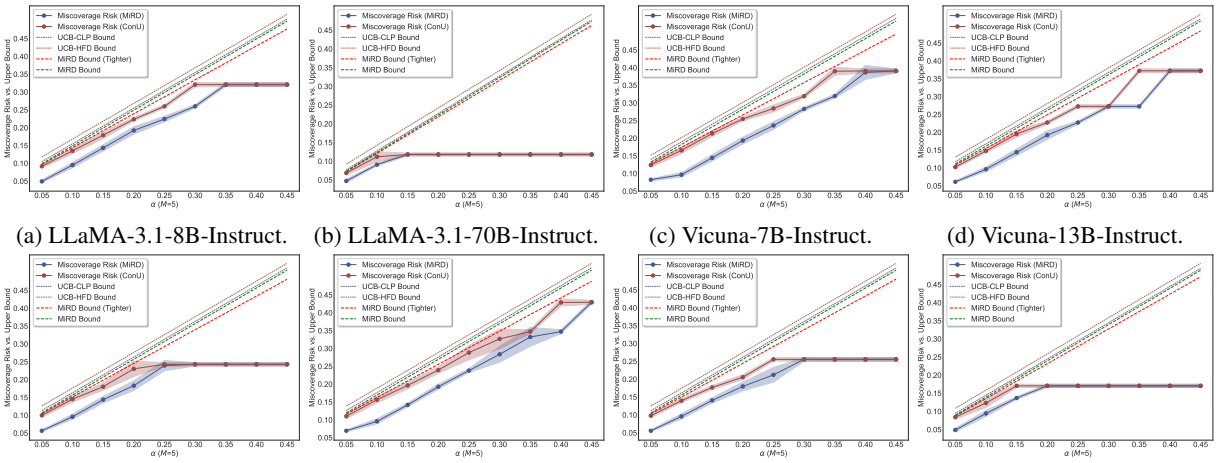
indicating that it is substantially tighter than PAC-style confidence bounds while still remaining valid. This is precisely the advantage of the expectation-level formulation in Proposition 1: it yields a non-vacuous and empirically sharp characterization of finite-sampling risk without introducing an additional (external) confidence parameter δ.

Tables 1 and 2 further demonstrate that this validity is robust to the admissibility criterion. Under sentence similarity, the bound continues to closely track the empirical sampling-failure risk across a range of semantic thresholds; under bi-entailment, the same behavior is observed for both 8B and 70B backbone models. These results suggest that the first-stage validity of MiRD is stable not only across model families and sampling budgets, but also across different notions of answer correctness.

4.3 Conditional Selection Risk Control under Full-Calibration Thresholding

We next evaluate the statistical validity of Stage II, fixing $M = 5$ on TriviaQA. As shown in Figure 3, MiRD consistently controls the conditional selection risk below the target α across all eight LLMs, validating the Stage-II guarantee in practice. Moreover, MiRD achieves uniformly lower test-time conditional miscoverage than ConU under the same risk level. This behavior is expected: by retaining the full calibration set, MiRD calibrates its threshold against a larger admission-correlated score distribution, which results in a weakly larger threshold than successful-only calibration and therefore a weakly larger prediction set on sampling-successful test examples. As discussed in Appendix A.3, this monotone enlargement preserves coverage and explains why MiRD is empirically more conservative than ConU in the selection stage.

However, lower conditional miscoverage alone does not fully characterize the utility of a set-valued predictor, because a larger threshold typically leads to a larger prediction set. This should not be viewed as a weakness by default. In set-valued prediction, the role of prediction-set size is not to be minimized uniformly, but to adapt to sample difficulty and reflect model uncertainty. A useful predictor should allocate larger sets to more ambiguous examples,



(a) LLaMA-3.1-8B-Instruct. (b) LLaMA-3.1-70B-Instruct. (c) Vicuna-7B-Instruct. (d) Vicuna-13B-Instruct.

(e) OpenChat-3.5 (7B). (f) Qwen2.5-3B-Instruct. (g) Qwen2.5-7B-Instruct. (h) Qwen2.5-14B-Instruct.

Figure 6: Overall miscoverage risk vs. upper bound at various risk levels on TriviaQA with eight LLMs ($M = 5$).

rather than collapsing to uniformly small sets.

To examine this behavior, we analyze deduplicated prediction set size using Vicuna-7B-Instruct on TriviaQA, stratifying sampling-successful test examples into easy and hard groups according to whether the top-1 candidate answer is admissible. As demonstrated in Figure 4, MiRD yields consistently larger and more stable prediction sets than ConU for both groups, avoiding the rapid shrinkage observed under successful-only calibration. More importantly, Figure 5 shows that MiRD preserves a larger adaptiveness gap between hard and easy examples across risk levels. This indicates that the additional set size introduced by MiRD is not merely uniform inflation; rather, it better retains difficulty-dependent variation and makes prediction-set size a more informative indicator of sample uncertainty.

Taken together, these results suggest that preserving full calibration-set integrity improves not only conditional risk control, but also the adaptiveness of the resulting set-valued predictions.

4.4 Overall Miscoverage Risk Control

We finally evaluate the end-to-end reliability. Since Stage I controls the marginal risk of finite-sampling failure, and Stage II controls conditional selection risk given sampling success, the resulting overall prediction set is expected to satisfy the population-level guarantees in Section 3.4 and Appendix A.4. Figure 6 confirms this empirically on TriviaQA: across all eight LLMs and all target risk levels considered, the test-time overall miscoverage of MiRD remains below both the conservative main-text bound and the tighter refinement in Appendix A.4.

Moreover, the tighter MiRD bound consistently tracks the empirical overall miscoverage at test time more closely than the conservative bound, confirming that the overlap correction in Appendix A.4 is meaningful in practice rather than merely algebraic. In contrast, UCB-CLP and UCB-HFD remain substantially looser, reflecting the conservativeness of PAC-style confidence bounds when propagated through the two-stage pipeline. Meanwhile, compared with ConU, MiRD achieves competitive or lower overall miscoverage while additionally accounting for finite-sampling failure, showing that preserving full calibration-set integrity improves end-to-end reliability rather than merely enlarging intermediate prediction sets.

Additional results on the CoQA and NQ datasets, and broader settings are provided in Appendix D.

5 Conclusion

We present MiRD, a two-stage framework for reliable set-valued prediction in open-ended QA. The central idea is to explicitly decompose overall miscoverage into two sources: marginal sampling failure and conditional selection failure. Based on this decomposition, MiRD first establishes a marginal upper bound in expectation for the finite-sampling risk, and further calibrates a conformalized selection threshold using the full calibration set through ground-truth-guided nonconformity estimates. Empirically, MiRD achieves valid risk control in both stages and for the overall prediction set across diverse LLM families, while producing tighter first-stage bounds than PAC-style confidence intervals and more adaptive prediction sets than successful-only calibration. These results demonstrate that preserving calibration-set integrity is beneficial not only for statistical validity, but also for the utility of set-valued prediction in open-ended generation.

Limitations

MiRD inherits several standard assumptions from conformal prediction and open-ended generation. First, its guarantees rely on the exchangeability between calibration and test examples, as in standard split conformal prediction. This assumption is appropriate for the marginal validity studied in this work, but substantial distribution shifts may require additional adaptation or recalibration mechanisms. Second, the selection-stage guarantee uses a score-level conditional compatibility assumption. We make this assumption explicit and provide a discussion in the appendix; developing weaker or empirically testable variants of this condition is an interesting direction for future work. Finally, MiRD studies risk control under a fixed finite sampling budget. This setting allows a clean decomposition of sampling failure and post-selection failure, while adaptive budget allocation could further improve the cost–risk trade-off in future deployments.

References

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. 2024. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*.

Anastasios N Angelopoulos and Stephen Bates. 2023. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. 2025. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

John J Cherian, Isaac Gibbs, and Emmanuel J Candès. 2024. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842.

Charles J Clopper and Egon S Pearson. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

Jesse C Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. 2024. Conformal prediction sets improve human decision making. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9439–9457.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Jinhao Duan, Xinyu Zhao, Zhuoxuan Zhang, Eunhye Grace Ko, Lily Boddy, Chenan Wang, Tianhao Li, Alexander Rasgon, Junyuan Hong, Min Kyung Lee, and 1 others. 2025. Guidellm: Exploring llm-guided conversation with applications in autobiography interviewing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5558–5588.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Yonatan Geifman and Ran El-Yaniv. 2019. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pages 2151–2159.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Wassily Hoeffding. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, and 1 others. 2024. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270.

Zijun Jia, Yuanchang Ye, Sen Jia, Yiyao Qian, Haoning Wang, Baojie Chen, Diyin Tang, Jinsong Yu, and Zhiyuan Wang. 2026. Balancerag: Joint risk calibration for cascaded retrieval-augmented generation. *arXiv preprint arXiv:2605.20084*.

Ying Jin and Emmanuel J Candès. 2023. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Ramneet Kaur, Colin Samplawski, Adam D. Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander Michael Berenbeim, John A. Pavlik, Nathaniel D. Bastian, and Susmit Jha. 2024. Addressing uncertainty in LLMs to enhance reliability in generative AI. In *Neurips Safe Generative AI Workshop 2024*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2305.18404*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Ye Li, Anqi Hu, Yuanchang Ye, Shiyan Tong, Zhiyuan Wang, and Bo Fu. 2026. Set-valued prediction for large language models with feasibility-aware coverage guarantees. *arXiv preprint arXiv:2603.22966*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*.

Harris Papadopoulos. 2008. *Inductive conformal prediction: Theory and application to neural networks*. INTECH Open Access Publisher Rijeka.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Yuan Sui, Yufei He, Zifeng Ding, and Bryan Hooi. 2025. Can knowledge graphs make large language models more trustworthy? an empirical study over open-ended question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12685–12701.

Binyu Tan, Zhiyuan Wang, Jinhao Duan, Kaidi Xu, Heng Tao Shen, Xiaoshuang Shi, and Fumin Shen. 2025. Conformal lesion segmentation for 3d medical images. *arXiv preprint arXiv:2510.17897*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024a. Openchat: Advancing open-source language models with mixed-quality data. In *The Twelfth International Conference on Learning Representations*.

Qingni Wang, Yue Fan, and Xin Eric Wang. 2026a. SAFER: Risk-constrained sample-then-filter in large language models. In *The Fourteenth International Conference on Learning Representations*.

Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. 2025a. Sample then identify: A general framework for risk control and assessment in multimodal large language models. In *The Thirteenth International Conference on Learning Representations*.

Sean Wang, Yicheng Jiang, Yuxin Tang, Lu Cheng, and Hanjie Chen. 2025b. Copu: Conformal prediction for uncertainty quantification in natural language generation. *arXiv preprint arXiv:2502.12601*.

Zhiyuan Wang, Tianlong Chen, Yue Zhang, Heng Tao Shen, Xiaoshuang Shi, Kaidi Xu, and 1 others. 2025c. Lec: Linear expectation constraints for false-discovery control in selective prediction and routing systems. *arXiv preprint arXiv:2512.01556*.

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

Zhiyuan Wang, Jinhao Duan, Qingni Wang, Xiaofeng Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. 2026b. Coin: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33764–33772.

Zhiyuan Wang, Jinhao Duan, Chenxi Yuan, Qingyu Chen, Tianlong Chen, Yue Zhang, Ren Wang, Xiaoshuang Shi, and Kaidi Xu. 2025d. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025e. SConU: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. 2025. Conformal prediction: A data perspective. *ACM computing surveys*, 58(2):1–37.

A Proofs and Discussions

A.1 Proof of Eq. (6)

By the law of total probability,

$$\begin{aligned}&\Pr(R_{N+1}(\lambda)=1)=\\&\Pr(R_{N+1}(\lambda)=1\mid Z_{N+1}=1)\Pr(Z_{N+1}=1)+\\&\Pr(R_{N+1}(\lambda)=1\mid Z_{N+1}=0)\Pr(Z_{N+1}=0).\end{aligned}$$

When $Z_{N+1}=1$, no admissible answer appears in the candidate set $\mathcal{G}_M(x_{N+1})$. Since $\mathcal{C}_\lambda(x_{N+1})\subseteq\mathcal{G}_M(x_{N+1})$, it follows that $R_{N+1}(\lambda)=1$ almost surely, so

$$\Pr(R_{N+1}(\lambda)=1\mid Z_{N+1}=1)=1.$$

This yields Eq. (6).

A.2 Proof of Proposition 1

Proof. By exchangeability, the random variables

$$Z_1(M),\ldots,Z_N(M),Z_{N+1}(M)$$

have the same marginal distribution. Hence,

$$\begin{aligned}\mathbb{E}[Z_{N+1}(M)]&=\frac{1}{N+1}\sum_{i=1}^{N+1}\mathbb{E}[Z_i(M)]\\&=\mathbb{E}\left[\frac{1}{N+1}\sum_{i=1}^{N+1}Z_i(M)\right].\end{aligned}\tag{23}$$

Using $Z_{N+1}(M)\le 1$, we obtain

$$\begin{aligned}&\frac{1}{N+1}\sum_{i=1}^{N+1}Z_i(M)\\&=\frac{1}{N+1}\sum_{i=1}^{N}Z_i(M)+\frac{1}{N+1}Z_{N+1}(M)\\&\le\frac{1}{N+1}\sum_{i=1}^{N}Z_i(M)+\frac{1}{N+1}.\end{aligned}$$

Taking expectations on both sides and using Eq. (8) gives

$$\mathbb{E}[Z_{N+1}(M)]\le\mathbb{E}\left[\frac{N}{N+1}\widehat{R}_N(M)+\frac{1}{N+1}\right].$$

Combining this with Eq. (9) proves the statistical validity of Proposition 1. □

A.3 Conditional Exchangeability Assumption

The second-stage guarantee in Section 3.3 requires that, conditional on $\{Z_{N+1}=0\}$, the augmented admission-correlated score sequence

$$\{u_1,\ldots,u_N,u_{N+1}\}$$

is exchangeable, a score-level assumption tailored to our construction and stronger than the exchangeability of the underlying data points alone.

The requirement for this assumption arises from an intrinsic asymmetry in our two-stage framework. In Stage I, the sampling failure is defined with respect to the finite candidate set generated by the model. In Stage II, however, conformalized selection is only meaningful on test data points for which sampling has already succeeded (i.e., $Z_{N+1} = 0$). By contrast, at calibration time, the ground-truth answer is always available, which allows us to define an admission-correlated score for *every* calibration data point, regardless of whether its sampled candidate set contains an admissible answer. Consequently, the calibration-side score distribution is constructed on the full calibration set, while the test-side score is only employed on the sampling-successful subset of the test distribution.

This asymmetry is inherent to the proposed risk decomposition, but rather a direct consequence of the risk decomposition in Eq. (6): the first-stage term accounts for whether an admissible answer appears at all, while the second-stage term only concerns whether such an answer is retained after conformalized filtering. Since Stage II is conditioned on $\{Z_{N+1} = 0\}$, its theoretical validity does not require the calibration and test candidate sets themselves to be symmetric in a pointwise sense. Instead, what is required is that the calibration-time ground truth scores provide a valid surrogate for the test-time admission-correlated reference (admissible) score under the same conditional regime.

More concretely, the calibration scores are defined as

$$u_i = U_G(y_i^* \mid x_i),$$

whereas the test score is defined using the lowest-uncertainty admissible sampled answer,

$$u_{N+1} = U_G(\hat{y}_{N+1}^{(\mathrm{ref})} \mid x_{N+1}),$$

where

$$\hat{y}_{N+1}^{(\mathrm{ref})} \in \operatorname*{arg\,min}_{\hat{y} \in G_M(x_{N+1}):\, A(y_{N+1}^*, \hat{y})=1} U_G(\hat{y} \mid x_{N+1}).$$

Therefore, exchangeability required in Section 3.3 is not implied automatically by the exchangeability of the raw data points (x_i, y_i^*). Rather, it is a compatibility assumption asserting that, conditional on sampling success, ground truth score distribution on calibration data is symmetric with reference admissible-score distribution on the future test data. Thus, the assumption formalizes the precise role of ground-truth answers: they are not injected into the candidate sets, nor do they alter the first-stage sampling event. Instead, they are used only to construct admission-correlated NSs on the calibration side, so as to retain all calibration examples while avoiding the exchangeability mismatch induced by augmentation-based designs. The validity of Stage II therefore hinges on score-level conditional compatibility, rather than candidate-set-level identity between calibration and test examples.

Then, we note that this assumption is empirically testable in a limited but informative sense. On calibration examples with $Z_i(M) = 0$, one may compare the ground truth score $U_G(y_i^* \mid x_i)$ against the corresponding admissible sampled-answer reference score

$$\min_{\hat{y} \in G_M(x_i):\, A(y_i^*, \hat{y})=1} U_G(\hat{y} \mid x_i),$$

to evaluate whether the two scores are well aligned in distribution or rank. Such analyses do not prove the assumption, but they offer evidence for the plausibility of using ground truth scores as surrogates for test-time admission-correlated sampled-answer scores.

A simpler interpretation based on nested conformal sets. For the specific nonconformity construction considered here, the Stage-II validity of MiRD also admits a simpler interpretation through threshold monotonicity. Previous methods such as ConU (Wang et al., 2024b) assume that both calibration and test sampling sets contain at least one admissible candidate. Under this assumption, the admission-correlated NSs on calibration and test examples are exchangeable, and the resulting conformal threshold is therefore statistically valid.

MiRD differs from this successful-only calibration view only in that it retains the full calibration set. Concretely, for calibration data points whose sampling sets already contain admissible answers, MiRD uses the same admission-correlated nonconformity scores as ConU; for calibration examples whose sampling sets fail, MiRD effectively adds additional maximal scores equal to 1 (let $U(\cdot) \leq 1$). Therefore, under the same target risk level α, the empirical score distribution used by MiRD is obtained by augmenting the successful-only score distribution with extra values at the upper endpoint. As a result, the calibrated quantile cannot decrease:

$$\hat{\lambda}_{\mathrm{MiRD}} \geq \hat{\lambda}_{\mathrm{ConU}}.$$

Since the conformal prediction sets are nested in the threshold,

$$\lambda_1 \le \lambda_2 \quad \Longrightarrow \quad C_{\lambda_1}(x) \subseteq C_{\lambda_2}(x),$$

we further have

$$C_{\hat{\lambda}_{\mathrm{ConU}}}(x) \subseteq C_{\hat{\lambda}_{\mathrm{MiRD}}}(x) \qquad \text{for every } x.$$

Hence, whenever the successful-only conformal predictor achieves coverage, the MiRD prediction set, being a superset under the same α, also preserves that coverage. In this sense, for the present nonconformity design, MiRD can be viewed as a monotone enlargement of the successful-only conformal predictor, which offers an intuitive explanation for its Stage-II statistical validity.

This interpretation is specific to the current nonconformity construction, where failed calibration examples contribute maximal scores, and should be understood as a complementary justification rather than a replacement for the more general score-level conditional exchangeability formulation above.

A.4 A Tighter Population-Level Miscoverage Bound

In the main text, we present a conservative marginal guarantee

$$\Pr\Big(R_{N+1}(\hat{\lambda}) = 1\Big) \le \alpha + \mathbb{E}\Big[\widetilde{\beta}(M)\Big], \tag{24}$$

where

$$\widetilde{\beta}(M) = \frac{N}{N+1}\widehat{R}_N(M) + \frac{1}{N+1}$$

is the exchangeability-corrected empirical proxy introduced in Section 3.2. This bound is simple and sufficient for our main result, but it is conservative because it upper-bounds the second-stage contribution by α ($\Pr(Z_{N+1}(M) = 0) \le 1$) without exploiting the fact that selection failure can only occur when finite sampling has already succeeded.

To sharpen this characterization, let

$$\begin{aligned} p_{\mathrm{fail}}(M) &:= \Pr(Z_{N+1}(M) = 1)\,, \\ p_{\mathrm{sel}} &:= \Pr\Big(R_{N+1}(\hat{\lambda}) = 1 \mid Z_{N+1}(M) = 0\Big)\,. \end{aligned} \tag{25}$$

Then the decomposition in Section 3.4 can be written as

$$\begin{aligned} \Pr\Big(R_{N+1}(\hat{\lambda}) = 1\Big) &= p_{\mathrm{fail}}(M) \\ &\quad + p_{\mathrm{sel}}\big(1 - p_{\mathrm{fail}}(M)\big). \end{aligned} \tag{26}$$

This formulation makes explicit that the second-stage risk is incurred only on the subset of test data for which the first-stage finite-sampling step has already produced at least one admissible answer.

Proposition 2 (A tighter population-level miscoverage bound)**.** *Under the assumptions of Sections 3.2–3.4,*

$$\begin{aligned} \Pr\Big(R_{N+1}(\hat{\lambda}) = 1\Big) &\le \alpha + (1-\alpha)p_{\mathrm{fail}}(M) \\ &\le \alpha + (1-\alpha)\mathbb{E}\Big[\widetilde{\beta}(M)\Big]\,. \end{aligned} \tag{27}$$

Proof. Starting from Eq. (26), we have

$$\Pr\Big(R_{N+1}(\hat{\lambda}) = 1\Big) = p_{\mathrm{fail}}(M) + p_{\mathrm{sel}}\big(1 - p_{\mathrm{fail}}(M)\big).$$

By Eq. (23) in the main text, $p_{\mathrm{sel}} \le \alpha$. Therefore,

$$\begin{aligned} \Pr\Big(R_{N+1}(\hat{\lambda}) = 1\Big) &\le p_{\mathrm{fail}}(M) + \alpha\big(1 - p_{\mathrm{fail}}(M)\big) \\ &= \alpha + (1-\alpha)p_{\mathrm{fail}}(M). \end{aligned}$$

Finally, Proposition 1 in the main text gives

$$p_{\mathrm{fail}}(M) \le \mathbb{E}\Big[\widetilde{\beta}(M)\Big]\,,$$

which yields the second inequality in Eq. (27). □

Compared with the conservative upper bound in Eq. (24), the improvement is

$$\begin{aligned} &\big(\alpha + \mathbb{E}[\widetilde{\beta}(M)]\big) - \big(\alpha + (1-\alpha)\mathbb{E}[\widetilde{\beta}(M)]\big) \\ &\qquad = \alpha\,\mathbb{E}[\widetilde{\beta}(M)]. \end{aligned}$$

Hence, the gap is exactly the overlap that is double-counted by the loose upper bound. Importantly, this tightening is obtained by coupling the two error terms through the exact decomposition in Eq. (26); it does *not* follow from independently replacing $\Pr(Z_{N+1}(M) = 0)$ by $1 - \mathbb{E}[\widetilde{\beta}(M)]$.

As an immediate corollary, we obtain a tighter population-level coverage guarantee

$$\begin{aligned} &\Pr\big(\exists \hat{y} \in \mathcal{C}_{\hat{\lambda}}(x_{N+1}), A(y^*_{N+1}, \hat{y}) = 1\big) \\ &\ge 1 - \alpha - (1-\alpha)\mathbb{E}\Big[\widetilde{\beta}(M)\Big]\,. \end{aligned} \tag{28}$$

Eq. (27) is viewed as a tighter population-level refinement of the main-text guarantee, while Eq. (24) remains the simpler and more transparent statement in the main development.

B Additional Related Work

Necessity of Calibration Set Integrity. Previous SCP methods for open-ended QA typically discard calibration data points whose sampling sets fail to cover admissible candidates (Wang et al., 2024b, 2025a; Kaur et al., 2024). Under the i.i.d. assumption, however, this pruning contracts the empirical uncertainty distribution and restricts conformal validity to a smaller subset of test distribution (Wang et al., 2025e). Moreover, both TRON (Wang et al., 2025a) and SConU (Wang et al., 2025e) suggest, through the calibration of sampling budget, that if every calibration instance contains admissible answers within the sampling set, then test points are guaranteed to produce admissible candidates via finite sampling. Such a premise conflicts with the model-agnostic nature of SCP (Angelopoulos and Bates, 2023): for any deployed generative model, it is generally impossible to ensure that every input query is answerable or that an admissible answer lies within the model's finite-sample output space.

Uncertainty Estimation for LLM Generation. Uncertainty estimation for free-form generation has been widely investigated as a way to detect hallucinations and unreliable model responses. Early approaches often exploit sampling-based consistency: if multiple stochastic generations are semantically consistent, the model is considered more reliable; otherwise, disagreement among generations can indicate uncertainty (Manakul et al., 2023). Because natural language admits many surface forms with the same meaning, later work proposes semantic-level uncertainty measures, such as semantic entropy (Kuhn et al., 2023), relevance-aware token or sentence weighting (Duan et al., 2024), and word-sequence-level uncertainty (Wang et al., 2025d), to better capture uncertainty in free-form answers. These methods provide useful uncertainty scores for open-ended QA, but they are typically heuristic and do not by themselves provide finite-sample risk guarantees. MiRD is complementary to this line of work: it can use such uncertainty measures as candidate scores, while calibrating them into risk-controlled set-valued predictions through the proposed miscoverage decomposition.

Risk-Controlling Prediction and Conformal Generation. Beyond SCP, risk-controlling prediction frameworks calibrate prediction sets to satisfy user-specified loss constraints under distribution-free or finite-sample guarantees (Bates et al., 2021; Jin and Candès, 2023; Angelopoulos et al., 2025). Recent work has also adapted conformal ideas to language generation, including conformal language modeling, conformal factuality guarantees, and surveys of SCP for NLP (Quach et al., 2024; Mohri and Hashimoto, 2024; Campos et al., 2024; Li et al., 2026). These studies indicate the promise of conformal methods for reliable generation, but many of them either focus on controlling the final generated output, calibrating a sampling or rejection rule, or constructing guarantees without explicitly separating the risk caused by finite sampling from the risk caused by post-hoc filtering. MiRD differs by decomposing overall miscoverage into marginal sampling failure and conditional selection failure, which makes the role of finite sampling explicit and enables full-calibration conformalized selection.

Selective Prediction and Abstention. Selective prediction studies the problem of abstaining from uncertain predictions to trade coverage for reliability (Geifman and El-Yaniv, 2017, 2019). This idea has also influenced LLM reliability, where systems may abstain, defer, or filter outputs when uncertainty is high (Jia et al., 2026; Wang et al., 2026b, 2025c). However, selective prediction usually focuses on deciding whether to output a prediction, while MiRD focuses on producing a set-valued answer for open-ended QA. Moreover, MiRD distinguishes two different reasons why a set-valued predictor may fail: the admissible answer may never appear in the sampled candidate pool, or it may appear but be removed by the selection rule. This distinction is essential in open-ended generation, where the candidate space is not fixed and must be approximated by finite stochastic sampling.

C Details of Experimental Setup

Details of utilized Datasets. TriviaQA (Joshi et al., 2017) is a reading comprehension benchmark containing over 650K samples, which are authored by trivia enthusiasts and independently gathered evidence documents. CoQA (Reddy et al., 2019) is a large-scale dataset for building Conversational QA systems, with 127k questions with free-form answers, and each question is equipped with contextual information. Following Lin et al. (2024), we use the validation split of the `rc.nocontext` subset of TriviaQA, the development split of CoQA. We employ 8,000 question-answer pairs of TriviaQA (total 9,960 de-duplicated questions), and the full 7,983 questions of CoQA. For each sample, we adopt a one-shot prompt, as illustrated as follows:

```
### System:
This is a bot that correctly answers questions.

### User:
In 1968, who did radical feminist Valerie Solanas shoot and wound as he entered his New York studio?
### Assistant:
Andy Warhol

### User:
Who was the man behind The Chipmunks?
### Assistant:
```

```
### System:
This is a bot that correctly answers questions.
Once upon a time, in a barn near a farm house, there lived a little white kitten named Cotton. Cotton lived high up in a nice warm place above the barn ...
### User:
What color was Cotton?
### Assistant:
white

### User:
Where did she live?
### Assistant:
```

NQ (Kwiatkowski et al., 2019) is a more realistic and challenging closed-book dataset, containing questions from real users, and it requires QA systems to read and comprehend an entire Wikipedia article that may or may not contain the answer to the question. We utilize 2,000 QA samples from the validation split with a total of 3,610 questions. The prompt adopted in NQ is illustrated as follows:

```
### System:
This is a bot that correctly answers questions.
### User:
who breaks a tie in the us senate?
### Assistant:
The Vice President of the United States

### User:
which domain of life are humans members of?
### Assistant:
```

Details of Correctness Criteria. Following previous research (Duan et al., 2024; Wang et al., 2024b, 2025d, 2026a), similarity is measured by the stsb-roberta-large (Reimers and Gurevych, 2019) model. We set the threshold as 0.6 by default, i.e., candidate answers having above 0.6 semantic similarities with the ground truth are deemed correct. For the bi-entailment criterion, we leverage a natural language inference (NLI) model with DeBERTa-v3-large-mnli-fever-anli-ling-wanli (He et al., 2023) as the backbone. Here, given a QA pair, the NLI classifier predicts scores (logits) for three distinct classes: entailment, neutral, and contradiction. See Section 4.1 in Lin et al. (2024) for more details. If and only if neither direction is labeled as contradiction and one direction is entailment, we consider the two QA instances to be equivalent.

Hyperparameters. We fix the calibration-test split ratio at 0.5 by default. Following prior work (Duan et al., 2024; Wang et al., 2024b, 2025c), we use multinominal sampling for candidates with the generation `temperature` of 1.0 and `top-p` of 0.9.

D Additional Experimental Results

Robustness of conditional selection risk across sampling budgets. Figure 7 further evaluates the conditional selection risk of MiRD under different sampling budgets on TriviaQA, using Vicuna-7B-Instruct as the backbone. Across all sampling budgets from $M = 4$ to $M = 20$ and all target risk levels, MiRD consistently keeps the empirical conditional selection risk below the target level α. This confirms that the Stage-II calibration remains valid not only at the default sampling budget used in the main experiments, but also across a broad range of finite-sampling regimes. As α increases, the empirical risk also increases smoothly, reflecting the expected coverage–set-size trade-off induced by conformal thresholding.

Compared with ConU, MiRD generally yields lower or comparable conditional selection risk across sampling budgets. This is consistent with our theoretical interpretation: by retaining sampling-failed calibration examples rather than discarding them, MiRD calibrates a weakly more conservative threshold over the full calibration distribution. The results show that this conservativeness is stable with respect to the sampling budget, suggesting that the benefit of full-calibration thresholding is not tied to a particular choice of M.

Overall miscoverage control under a larger sampling budget. Figure 8 reports the end-to-end overall miscoverage risk on TriviaQA when the sampling budget is increased to $M = 20$. Across all eight LLMs, the empirical overall miscoverage of MiRD remains below the theoretical upper bounds over the full range of target risk levels.

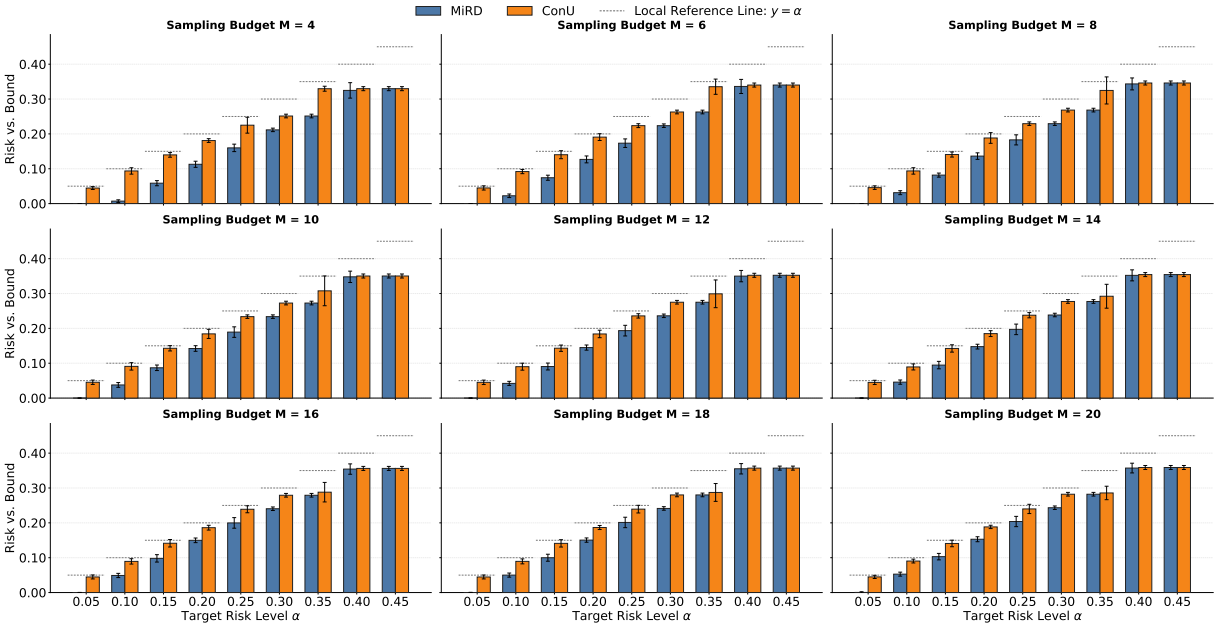
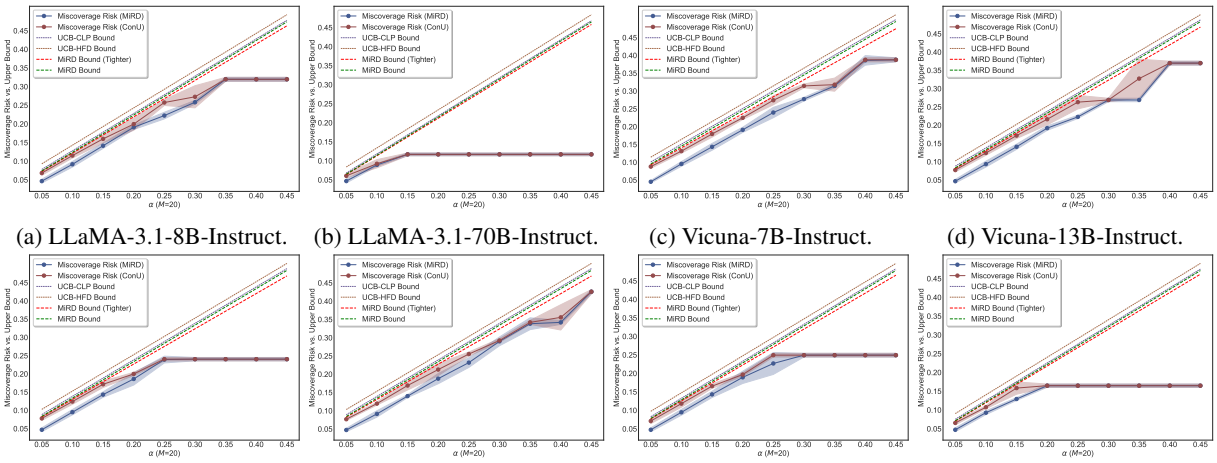


Figure 7: Conditional Selection Risk of MiRD vs. ConU across various sampling budgets and risk levels, using Vicuna-7B-Instruct on TriviaQA. Each subplot corresponds to one sampling budget. Within each α group, the dashed segment marks the target level α; bars lower than the dashed segment indicate the mean risk is below α.



(a) LLaMA-3.1-8B-Instruct. (b) LLaMA-3.1-70B-Instruct. (c) Vicuna-7B-Instruct. (d) Vicuna-13B-Instruct.

(e) OpenChat-3.5 (7B). (f) Qwen2.5-3B-Instruct. (g) Qwen2.5-7B-Instruct. (h) Qwen2.5-14B-Instruct.

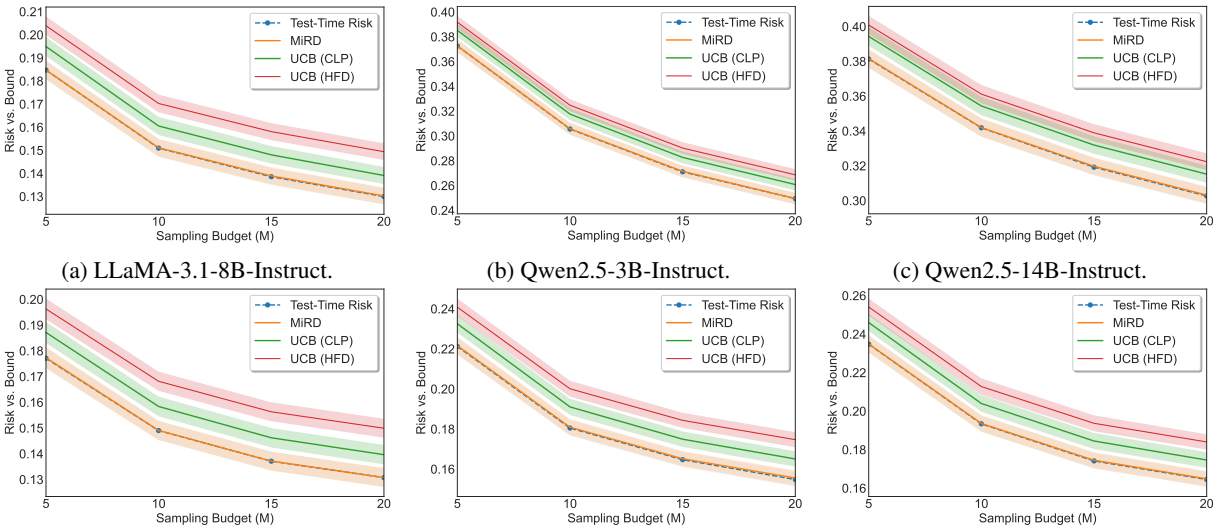
Figure 8: Overall miscoverage risk vs. upper bound at various risk levels (α) on TriviaQA with eight LLMs. Here, we set the sampling budget to 20.

Moreover, the tighter MiRD bound tracks the empirical miscoverage more closely than the conservative main-text bound, while both Clopper–Pearson and Hoeffding-style UCB bounds remain substantially looser. This further supports the usefulness of the refined decomposition bound in Appendix A.4.

The comparison with ConU also shows that MiRD maintains competitive or lower end-to-end miscoverage while explicitly accounting for finite-sampling failure. This is important because increasing the sampling budget reduces, but does not conceptually remove, the possibility of sampling failure in open-ended QA. Therefore, even under a relatively large sampling budget, explicitly decomposing overall miscoverage into sampling failure and conditional selection failure remains beneficial for reliable set-valued prediction.

We further evaluate MiRD on CoQA, a context-supported open-ended QA benchmark. Compared with TriviaQA, CoQA provides supporting passages and therefore represents a complementary setting where answer generation is conditioned on explicit context rather than relying purely on parametric knowledge. The results show that the proposed miscoverage decomposition remains effective in this setting.

Sampling-failure risk control on CoQA. Figure 9 illustrates the sampling-failure risk and its upper



(a) LLaMA-3.1-8B-Instruct. (b) Qwen2.5-3B-Instruct. (c) Qwen2.5-14B-Instruct.

(d) OpenChat-3.5 (7B). (e) Vicuna-7B-V1.5. (f) Vicuna-13B-V1.5.

Figure 9: Sampling risk vs. three upper bounds at various sampling budgets (M) on CoQA with six LLMs.

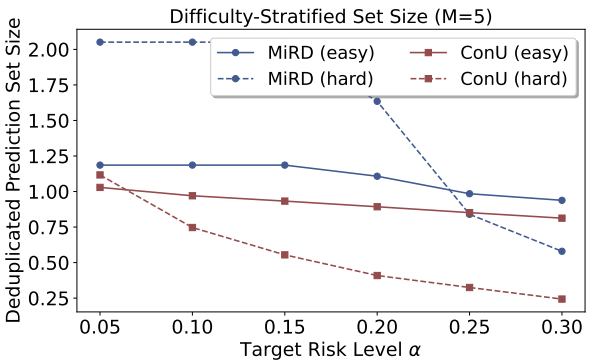


Figure 10: Difficulty-stratified deduplicated prediction set size on CoQA with LLaMA-3.1-8B-Instruct. Test samples are partitioned into easy and hard groups according to whether the top-1 candidate is admissible.

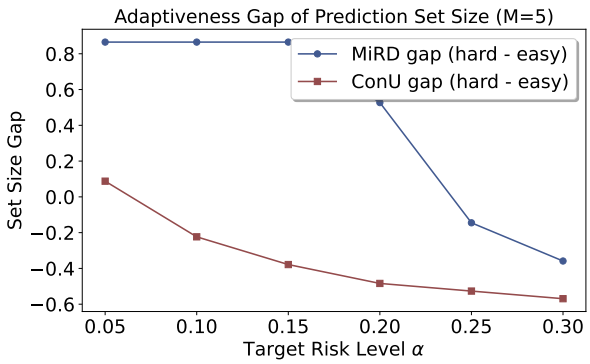


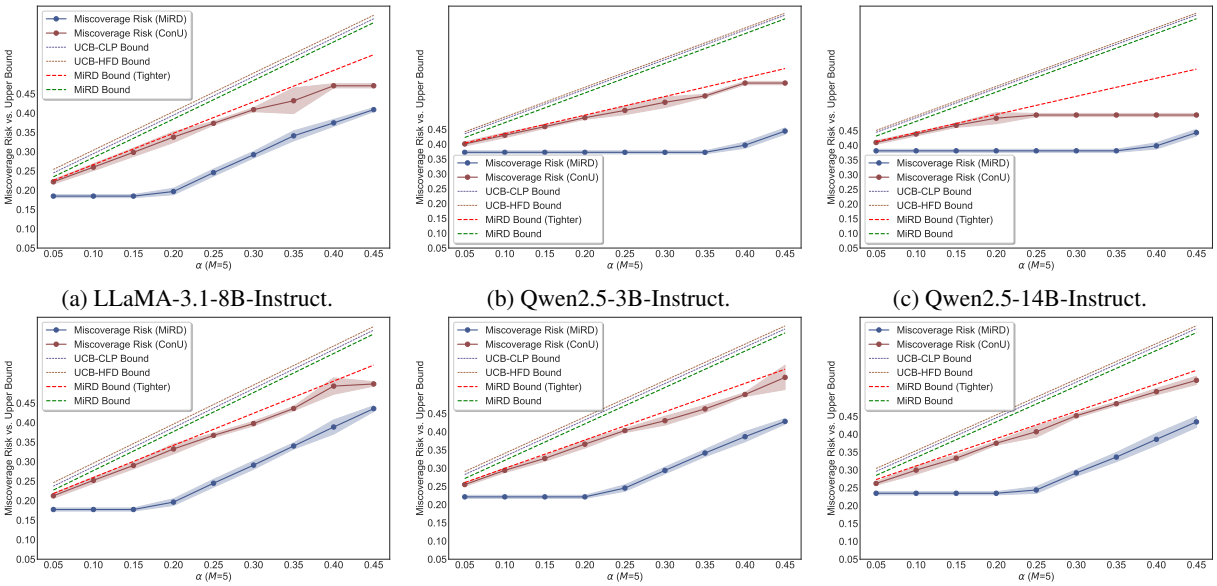
Figure 11: Adaptiveness gap of prediction set size on CoQA with LLaMA-3.1-8B-Instruct, defined as the difference between the average deduplicated prediction set sizes of hard and easy examples.

bounds on CoQA across six LLMs and multiple sampling budgets. Across all models and budgets considered, the MiRD bound consistently stays above the empirical test-time sampling-failure risk, validating the expectation-level marginal guarantee of Stage I. At the same time, the MiRD bound is uniformly tighter than both Clopper–Pearson and Hoeffding-style UCB bounds. This confirms that the advantage of the proposed expectation-level formulation is not specific to TriviaQA: even in a context-supported QA setting, MiRD provides a non-vacuous and empirically sharp characterization of finite-sampling failure.

As expected, the sampling-failure risk decreases as the sampling budget M increases. However, the risk remains non-negligible even at larger budgets for several backbones, showing that finite sampling cannot be ignored simply by increasing M. This supports the central motivation of MiRD: reliable open-ended set-valued prediction should explicitly account for the possibility that no admissible answer is sampled, rather than treating reliability only as a post-hoc filtering problem.

Prediction-set adaptiveness on CoQA. Figures 10 and 11 analyze the deduplicated prediction set size on CoQA using LLaMA-3.1-8B-Instruct. We again divide sampling-successful test examples into easy and hard groups according to whether the top-1 candidate is admissible. MiRD produces larger prediction sets for hard examples than for easy examples over a broad range of risk levels, indicating that its set size reflects example difficulty. By contrast, ConU tends to shrink prediction sets more aggressively, especially on hard examples.

The adaptiveness-gap plot further highlights this difference. MiRD maintains a positive hard–easy



(a) LLaMA-3.1-8B-Instruct. (b) Qwen2.5-3B-Instruct. (c) Qwen2.5-14B-Instruct.

(d) OpenChat-3.5 (7B). (e) Vicuna-7B-V1.5. (f) Vicuna-13B-V1.5.

Figure 12: Overall miscoverage risk vs. upper bound at various risk levels on CoQA with six LLMs ($M = 5$).

set-size gap at stricter risk levels, while ConU exhibits a much smaller or negative gap across most risk levels. This suggests that successful-only calibration may distort the relationship between set size and sample difficulty, whereas MiRD better preserves uncertainty-dependent variation by retaining the full calibration set. Thus, the additional prediction-set size introduced by MiRD is not merely uniform inflation; it is more aligned with the ambiguity of the test example.

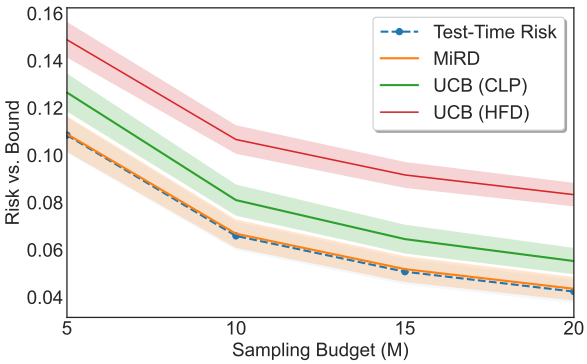
Overall miscoverage risk control on CoQA. Figure 12 evaluates the end-to-end overall miscoverage risk on CoQA with $M = 5$. Across all six LLMs, the empirical overall miscoverage of MiRD remains below the proposed upper bounds over the full range of target risk levels. Moreover, the tighter MiRD bound tracks the empirical risk more closely than the conservative bound, while the PAC-style UCB bounds remain substantially looser. This again confirms the practical value of the refined decomposition bound.

Compared with ConU, MiRD achieves competitive or lower overall miscoverage while explicitly accounting for finite-sampling failure. This result is particularly important on CoQA because the presence of supporting context does not eliminate sampling uncertainty: a model may still fail to generate an admissible answer within a finite budget, or may generate one but remove it during filtering. The CoQA results therefore reinforce the main conclusion that reliable open-ended QA requires separating finite-sampling failure from conditional post-selection failure, rather than collapsing both into a single filtering-stage error.

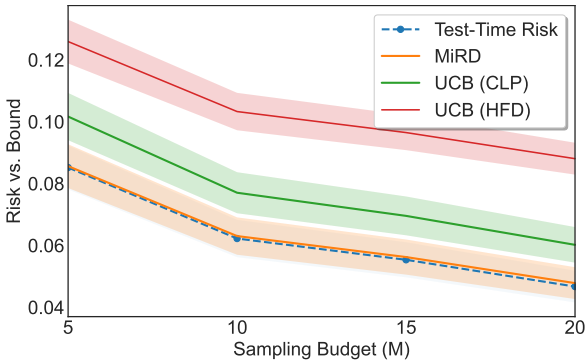
We finally evaluate MiRD on Natural Questions (NQ), a more challenging open-domain QA benchmark with real user questions. Compared with TriviaQA and CoQA, NQ poses a harder generation setting, making it a useful stress test for whether the proposed decomposition remains valid when finite sampling is more likely to fail. We report results under both sentence-similarity and bi-entailment correctness criteria.

Sampling-failure risk control on NQ. Figure 13 reports the sampling-failure risk and its upper bounds on NQ using two representative LLMs. Across both models, both correctness criteria, and all sampling budgets considered, the MiRD bound remains above the empirical test-time sampling-failure risk. This validates the expectation-level marginal guarantee of Stage I in a substantially more challenging QA setting. Moreover, MiRD continues to provide tighter bounds than both Clopper–Pearson and Hoeffding-style UCBs, showing that the advantage of the proposed marginal formulation is robust beyond the TriviaQA and CoQA settings.

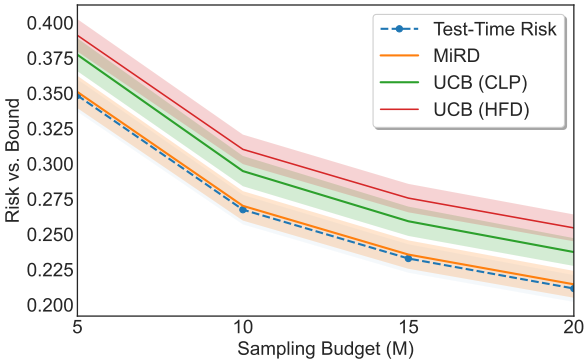
The results also reveal that the choice of admissibility criterion substantially affects the magnitude of sampling failure. Under the stricter bi-entailment criterion, the empirical sampling-failure risk is noticeably higher than under sentence similarity, especially for weaker backbones. This observation further supports the need to treat finite-



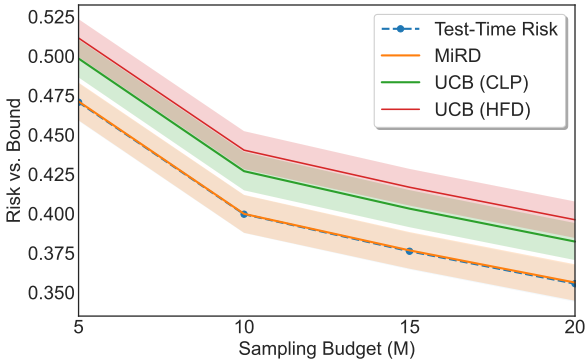
(a) LLaMA-3.1-8B-Instruct (similarity).



(b) Qwen2.5-3B-Instruct (similarity).



(c) LLaMA-3.1-8B-Instruct (entailment).



(d) Qwen2.5-3B-Instruct (entailment).

Figure 13: Sampling risk vs. three upper bounds at various sampling budgets (M) on NQ with two LLMs.

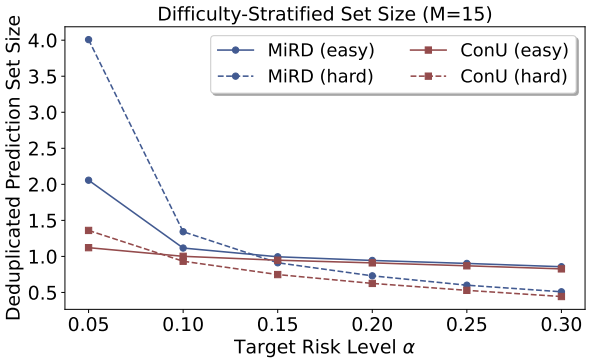


Figure 14: Difficulty-stratified deduplicated prediction set size on NQ with LLaMA-3.1-8B-Instruct. Test samples are partitioned into easy and hard groups according to whether the top-1 candidate is admissible.

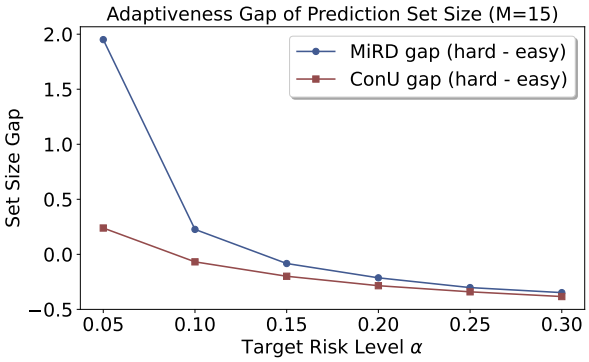
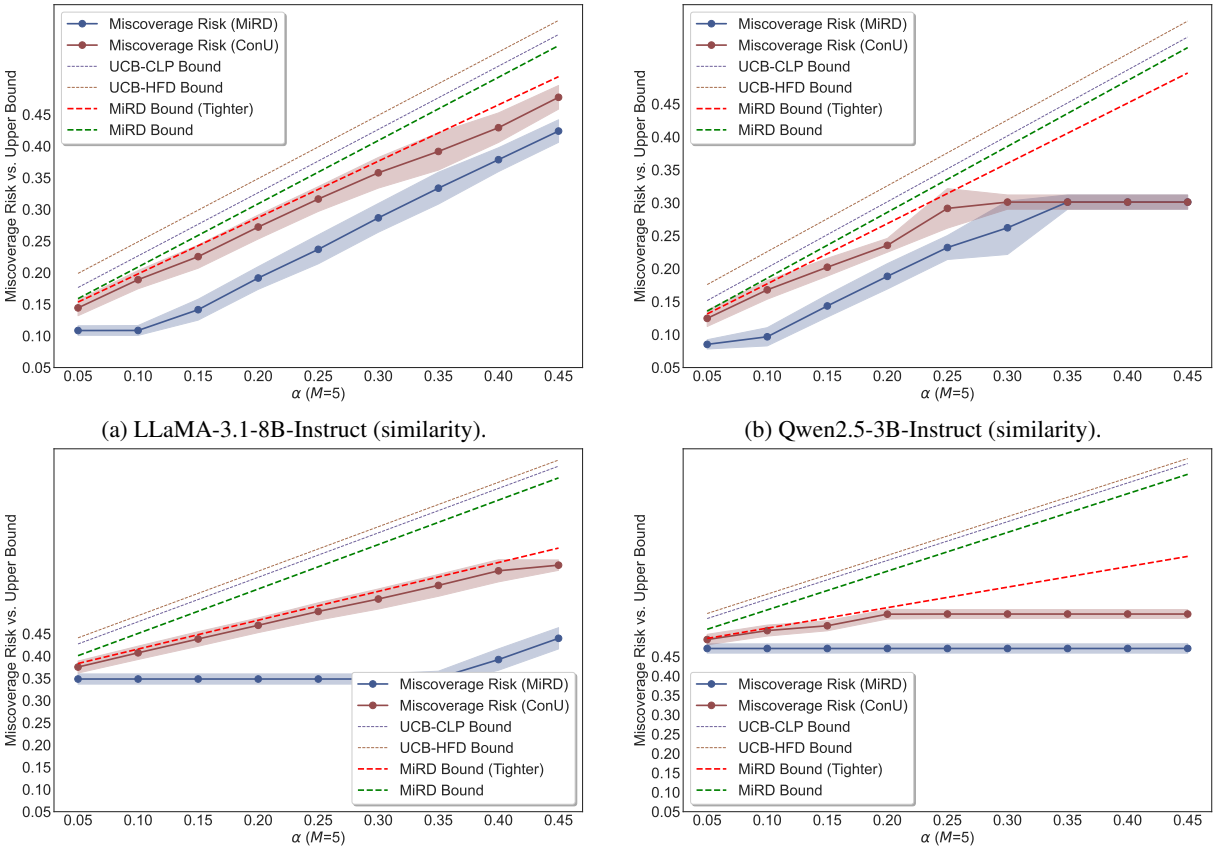


Figure 15: Adaptiveness gap of prediction set size on NQ with LLaMA-3.1-8B-Instruct, defined as the difference between the average deduplicated prediction set sizes of hard and easy examples.

sampling failure as an explicit component of the reliability objective: when correctness is evaluated more conservatively, simply increasing the sampling budget may reduce but does not eliminate the risk that no admissible answer appears in the candidate set.

Prediction-set adaptiveness on NQ. Figures 14 and 15 analyze the deduplicated prediction set size on NQ using LLaMA-3.1-8B-Instruct. As in the main experiments, we divide sampling-successful test examples into easy and hard groups according to whether the top-1 candidate is admissible. MiRD assigns larger prediction sets to hard examples than to easy examples across most risk levels, indicating that its prediction sets remain sensitive to example difficulty even on NQ.

The adaptiveness-gap plot further confirms this behavior. MiRD maintains a consistently larger hard–easy set-size gap than ConU, while ConU exhibits a much smaller gap and may even under-allocate set size to hard examples. This suggests that successful-only calibration can weaken the relationship between set size and sample difficulty, whereas MiRD better preserves uncertainty-



(a) LLaMA-3.1-8B-Instruct (similarity).

(b) Qwen2.5-3B-Instruct (similarity).

(c) LLaMA-3.1-8B-Instruct (entailment).

(d) Qwen2.5-3B-Instruct (entailment).

Figure 16: Overall miscoverage risk vs. upper bound at various risk levels on NQ with two LLMs ($M = 5$).

dependent variation by calibrating over the full calibration distribution.

Overall miscoverage risk control on NQ. Figure 16 evaluates the end-to-end overall miscoverage risk on NQ. Across both LLMs and both correctness criteria, the empirical overall miscoverage of MiRD remains below the proposed upper bounds over the range of target risk levels. The tighter MiRD bound again tracks the empirical miscoverage more closely than the conservative bound, while PAC-style UCB bounds remain looser. These results confirm that the refined decomposition bound remains useful even when the underlying sampling-failure risk is relatively high.

Compared with ConU, MiRD achieves competitive or lower overall miscoverage while explicitly accounting for finite-sampling failure. This is particularly important on NQ, where the harder question distribution and stricter admissibility criteria make sampling failure more pronounced. Together with the TriviaQA and CoQA results, the NQ experiments demonstrate that MiRD provides consistent risk control across knowledge-intensive, context-supported, and more challenging open-domain QA settings.

E Use of AI Assistants

We used AI assistants (e.g., ChatGPT) to assist with grammar and proofreading in our paper writing.