Conformal Non-Coverage Risk Control (CNCRC): Risk-Centric Guarantees for Predictive Safety in High-Stakes Settings

Che Shen [1] **Junwei Su** [2] **Lingpeng Kong** [1] **Chuan Wu** [1]

Abstract

Standard Conformal Prediction (CP) guarantees that prediction sets contain the true label with high probability, but it is *cost-blind*, treating all errors as equally important—a critical limitation in high-stakes domains. We introduce *Conformal Non-Coverage Risk Control (CNCRC)*, a framework that replaces coverage frequency with direct risk control. CNCRC guarantees an upper bound on catastrophic *non-coverage risk* while actively reducing *ambiguity risk*, providing prediction sets that are both safe and usable. This is achieved through a principled decomposition of decision risk and the design of risk-weighted nonconformity scores that balance robustness with efficiency. Experiments show that CNCRC reliably satisfies strict risk constraints in adversarial settings and outperforms all baselines on a large-scale clinical benchmark. By offering practitioners a choice between maximum robustness and maximum efficiency, CNCRC provides a practical and theoretically grounded toolkit for deploying genuinely risk-aware machine learning systems in safety-critical applications.

1. Introduction

A fundamental challenge for the reliable deployment of machine learning models is the quantification of their predictive uncertainty (Huang et al., 2025). While recent deep learning models can achieve high accuracy, they often produce point predictions without a formal measure of confidence, making it difficult to distinguish a confident guess from a borderline one (Amodei et al., 2016; Guo et al., 2017). Among the various methods for Uncertainty Quantification (UQ) (Lakshminarayanan et al., 2017; MacKay, 1992; Ye et al., 2024), Conformal Prediction (CP) has emerged as a particularly compelling framework due to its elegant theoretical properties (Angelopoulos & Bates, 2021). At its core, CP generates a *prediction set*—a small collection of plausible labels—that is mathematically guaranteed to contain the true outcome with a user-specified probability (e.g., 95%) (Vovk et al., 2005). For example, an image classifier augmented by CP would not just output 'cat'; it can provide a set like {'cat', 'lynx'} with a 95% guarantee that the true class is present in the set, transparently communicating the model's ambiguity to the end-user. This rigorous statistical guarantee holds without making any assumptions about the data's distribution, a highly desirable property known as being distribution-free (Vovk et al., 2005; Angelopoulos & Bates, 2021). Furthermore, the framework is *model-agnostic*: it functions as a lightweight, post-processing step that can be applied to any pre-trained model, from simple classifiers to complex Large Language Models (LLMs), without altering their internal architecture (Angelopoulos & Bates, 2021; Wang et al., 2024). This combination of a rigorous, distribution-free guarantee and universal applicability makes CP a powerful and versatile tool.

The Failure of Conformal Prediction in High-Stakes Domains. Despite its powerful and general guarantees, classical CP fails in high-stakes domains due to a critical flaw: its guarantee is fundamentally *cost-blind* (Wilder et al., 2019). The standard procedure controls the long-run frequency of coverage, that is, the proportion of test instances in which the prediction set contains the true label, but it treats all errors as equally consequential, ignoring the often asymmetric costs of real-world mistakes (Elkan, 2001; Elmachtoub & Grigas, 2022). This oversight can lead to catastrophic failures. To make this concrete, consider a clinical scenario with a rare but lethal condition—a "poison class". A model might predict the class probabilities for a given patient as follows: a common benign condition at 60%, another at 39.9%, and the poison class at a mere 0.1%. A risk-blind classical CP aims to form a confidence set by capturing a certain amount of probability mass, typically by selecting the most likely candidates. In this case, such a

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, China [2]School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, China. Correspondence to: Chuan Wu <cwu@cs.hku.hk>, Junwei Su <junweisu.cs@gmail.com>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

procedure would form a set containing only the two benign conditions, as their combined probability already accounts for 99.9% of the distribution. The poison class, despite its critical importance to the patient's survival, is systematically excluded simply because its probability is low. This illustrates the core problem: any method that is blind to the asymmetric costs of errors and relies solely on a probabilistic criterion is fundamentally unsafe for high-stakes decisions (Elmachtoub & Grigas, 2022; Sadinle et al., 2019). Defining these asymmetric costs is therefore not an optional assumption, but a fundamental prerequisite for safety in such domains (Elkan, 2001). In the example above, the system's statistical "confidence" provides no true measure of clinical safety.

Limitation of Existing Approaches. The failure of classical CP to account for such cost asymmetries has motivated some recent studies to incorporate cost-awareness. One influential line of work is cost-aware CP, which seeks to generate sets that optimize a downstream utility function (Cortes-Gomez et al., 2024). While this can improve average-case utility, its reliance on expectations is precisely its undoing in high-stakes settings, as it offers no hard guarantee against specific, high-cost failures. An average-case metric is easily dominated by the outcomes of frequent, low-cost events, meaning the immense cost of a single, rare event—such as failing to identify the "poison class"—can be effectively ignored or "averaged out" in the optimization. A more formal approach is Conformal Risk Control (CRC), which provides a framework to control a user-defined expected loss (Angelopoulos et al., 2022). However, CRC's guarantees come with practical trade-offs. To formally control the expected loss for a general loss function in a distribution-free manner, the framework must be inherently conservative to account for worst-case scenarios. In practice, this conservatism often forces the procedure to generate unmanageably large prediction sets which, by including numerous distractors, create high ambiguity and are of little use to a decision-maker (Lu et al., 2022). Consequently, a critical gap remains for a framework that can move beyond average-case performance and instead provide direct, rigorous control over catastrophic risks, while simultaneously ensuring the practical utility of its outputs.

Our Approach. To resolve the tension between CP's statistical promise and its practical failure in cost-sensitive settings, we propose *Conformal Non-Coverage Risk Control (CNCRC)*, a framework that fundamentally redefines the conformal guarantee. It shifts the objective from controlling *error frequency*, which treats all mistakes as equal, to directly controlling the *decision risk*—the expected real-world consequence of a prediction, where each potential error is weighted by its asymmetric cost. Our central conceptual breakthrough is the insight that this tension can be resolved through a principled decomposition of the monolithic notion of 'risk' into two distinct, actionable components. First, it treats *non-coverage risk*—the expected loss from a catastrophic failure when the true label is absent from the prediction set—as a hard constraint. Instead of targeting an average-case expectation, CNCRC provides a rigorous and numerically interpretable upper bound on this risk, a crucial feature for controlling the high-cost tail events characteristic of "poison class" scenarios. Second, the framework explicitly minimizes *ambiguity risk*, which refers to the potential harm caused by incorrect but plausible "distractor" labels remaining inside the prediction set, since a decision-maker may mistakenly select one of these high-risk options. These distractors are incorrect yet plausible options that are particularly dangerous when they are high-risk—for instance, a contraindicated drug appearing alongside the correct, safe therapy. By actively purging these hazardous options, CNCRC ensures the final output is not just statistically valid but also decision-aware and safe, preventing the set from becoming a "potential trap" for the end user. To achieve this dual objective, CNCRC replaces the cost-blind nonconformity score of classical CP with a family of novel, risk-weighted alternatives, which integrate asymmetric error costs directly into the conformal calibration process. Crucially, while providing these stronger, risk-based assurances, our framework is proven to retain the standard split-conformal marginal coverage guarantee, adding a new layer of risk control without sacrificing the foundational properties of CP.

Contributions. The main contributions of this paper are summarized as follows.

A Principled Reformulation of High-Stakes Uncertainty Quantification. We introduce a new perspective on conformal prediction that shifts the objective from controlling statistical error frequency to directly controlling decision risk. This is formalized through a novel risk decomposition into a hard constraint on non-coverage risk (R_{NC}) and a design objective of minimizing ambiguity risk ($\mathrm{AmbCost}$).

A Practical and Theoretically Grounded Algorithmic Framework. We propose CNCRC, which minimizes the proposed objective via novel risk-weighted nonconformity scores ($s_{\max}$, s_{sum}) that satisfy a Risk-Bounding Property. This property forms the theoretical bridge between conformal calibration and rigorous risk control, providing a theoretical guarantee to our method (Theorem 3.1).

Extensive Empirical Validation across Three Distinct Benchmarks. We evaluate CNCRC on a comprehensive suite of three benchmarks: (i) an *Adversarial Stress Test* to probe theoretical robustness against worst-case "poison" classes; (ii) a *Realistic Clinical Benchmark* to assess efficiency in high-dimensional settings; and (iii) a *Dynamic Risk Generalization* analysis to verify stability across vary-

ing risk budgets. Our results demonstrate that CNCRC is the only framework to strictly satisfy risk constraints in adversarial settings while consistently forming the Pareto frontier of the safety-efficiency trade-off in clinical applications.

2. Related Work

Our work builds upon the foundational framework of CP, seeking to improve upon existing methods for cost-awareness. We situate our contribution with respect to three key areas: the classical CP framework, utility-focused adaptations for cost-awareness, and the formal theory of CRC.

Conformal Prediction (CP). The goal of CP is not to provide a single "best" prediction, but rather a statistically reliable *prediction set* (Vovk et al., 2005; Sadinle et al., 2019), denoted $\mathcal{C}(\mathbf{x})$. The core promise of the framework is that this set is mathematically guaranteed to contain the true label, y_{true}, with a user-specified high probability (Vovk et al., 2005). In practice, this is most commonly achieved via Split CP, which requires holding out a separate *calibration set*, $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. The central component of the framework is a *nonconformity score*, $s(\mathbf{x}, y)$, a function that measures how poorly a label y conforms to an input $\mathbf{x}$ according to a pre-trained model (Papadopoulos et al., 2002; Angelopoulos & Bates, 2021). A threshold, q, is then determined by taking the $\lceil (n+1)(1-\alpha) \rceil / n$-th quantile of scores computed on the calibration set, where $s_i = s(\mathbf{x}_i, y_i)$ and α is the user-specified tolerable error rate. For a new test input $\mathbf{x}_{\text{new}}$, the prediction set is constructed as:

$$\mathcal{C}(\mathbf{x}_{\text{new}}) = \{y \in \mathcal{Y} \mid s(\mathbf{x}_{\text{new}}, y) \leq q\}. \tag{1}$$

Under the standard assumption that the data points are exchangeable, meaning that their joint distribution is invariant to permutations of the sample order, this procedure provides the rigorous marginal coverage guarantee that $P(y_{\text{true}} \in \mathcal{C}(\mathbf{x}_{\text{new}})) \geq 1 - \alpha$. This finite-sample, distribution-free validity fundamentally distinguishes CP from other Uncertainty Quantification (UQ) paradigms, such as Bayesian methods (MacKay, 1992; Lakshminarayanan et al., 2017), which typically rely on asymptotic or model-dependent assumptions. Modern advancements have also focused on making these sets adaptive to the difficulty of each prediction (Romano et al., 2020; Angelopoulos et al., 2020) or ensuring class-conditional coverage validity (Vovk et al., 2005; Sadinle et al., 2019). However, the entire mechanism, particularly the canonical nonconformity score $s(\mathbf{x}, y) = 1 - \hat{p}(y|\mathbf{x})$, where $\hat{p}(y|\mathbf{x})$ denotes the model's estimated probability of label y given input $\mathbf{x}$, relies solely on model probabilities. It is completely unaware of the real-world costs of different errors, a design choice that renders it not merely cost-blind but profoundly misaligned with the realities of high-stakes decision-making. This misalignment creates an urgent need for a new class of methods that treat asymmetric risks not as an afterthought, but as a core design principle.

Cost-Aware CP. To address the cost-blindness of classical CP, one influential line of work seeks to incorporate cost or utility information directly into the framework. A key example is Utility-Directed Conformal Prediction, which aims to generate prediction sets that optimize for an expected downstream utility (Cortes-Gomez et al., 2024). The central mechanism in this approach is to modify the nonconformity score by adding a cost-based penalty term. The score for a candidate label y can be expressed in a general form as:

$$s_{\text{cost-aware}}(\mathbf{x}, y) = s_{\text{base}}(\mathbf{x}, y) + \lambda \cdot \text{cost}(y), \tag{2}$$

where $s_{\text{base}}(\mathbf{x}, y)$ is a standard, probability-based score (e.g., $1 - \hat{p}(y|\mathbf{x})$), $\text{cost}(y)$ represents the disutility associated with label y, and λ is a hyperparameter that controls the strength of the cost penalty. While this method can improve the average utility of the resulting sets, its reliance on a tunable hyperparameter λ makes the procedure heuristic in nature. More fundamentally, optimizing expected utility "averages out" rare, high-cost events. Unlike these methods, CNCRC enforces a hard constraint ($R_{NC} \leq R_0$) on non-coverage risk, ensuring safety against catastrophic failure modes even when they are rare.

Conformal Risk Control (CRC) and Variants. A more rigorous approach to cost-awareness is *Conformal Risk Control (CRC)*, a framework that guarantees the expected value of a loss function remains below a desired level (Angelopoulos et al., 2022). This paradigm encompasses several related extensions, such as *Learn-then-test (LTT)* (Angelopoulos et al., 2025), which frames risk control as a multiple hypothesis testing problem, and *Conformal Tail Risk Control* (Chen et al., 2025), which targets heavy-tailed risks in language generation. While theoretically powerful, this entire class of generic risk-control methods faces a critical efficiency bottleneck when applied to cost-sensitive multi-class classification. The fundamental limitation is that these frameworks primarily focus on *identifying a safe threshold* for a given nonconformity score, rather than *reshaping the score itself*. Consequently, when the underlying score remains probability-centric (prioritizing high-probability but high-risk candidates), these methods are forced to correct for the risk by aggressively lowering the threshold. This conservatism results in unmanageably large prediction sets that satisfy the risk bound but obscure the decision-making process with excessive ambiguity. By contrast, our work directly confronts this limitation by redesigning the nonconformity score, enabling risk control without the efficiency penalty inherent to generic threshold-based frameworks.

3. Methodology

3.1. Problem Formulation

Preliminaries. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{y_1, \ldots, y_K\}$ be a discrete label space. For any input $\mathbf{x} \in \mathcal{X}$, a base predictor F produces a class-probability distribution $P(\cdot \mid \mathbf{x})$ over $\mathcal{Y}$. Our framework is model-agnostic, meaning F can be any model that outputs such a distribution (Angelopoulos & Bates, 2021; Romano et al., 2020). In fact, the only requirement is exchangeability of the data; no assumptions are made about model capacity or training. This independence underscores the generality and elegance of CNCRC. The key ingredient that allows our framework to move beyond the cost-blind nature of classical CP is a user-specified, asymmetric *cost function*, $\mathrm{Cost} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, where $\mathrm{Cost}(y_{\text{true}}, y_{\text{pred}})$ encodes the real-world consequence of predicting y_{pred} when the truth is y_{true}. We assume costs are bounded above by a finite constant $C_{\max} \in (0, \infty)$.[1]

Objectives: From Coverage to Risk Control. Our goal is to construct a prediction set $\mathcal{C}(\mathbf{x})$ that provides multi-faceted, risk-aware guarantees. We begin by preserving the foundational guarantee of split-conformal prediction: for a user-specified significance level $\alpha \in (0, 1)$, the set must satisfy marginal coverage, $\Pr\big(y_{\text{true}} \in \mathcal{C}(\mathbf{x})\big) \geq 1 - \alpha$. Beyond this, we introduce and aim to control two distinct, risk-centric objectives.

First, we control for catastrophic omissions via the *non-coverage risk*, defined as:

$$R_{\text{NC}} := \mathbb{E}\Big[\mathrm{Cost}\big(y_{\text{true}}, y_{\text{default}}\big) \cdot \mathbf{1}\{y_{\text{true}} \notin \mathcal{C}(\mathbf{x})\}\Big],$$

where y_{default} is a default fallback action, representing the system's pre-specified safe choice when the prediction set misses the true label (e.g., in a clinical setting, defaulting to "withhold treatment and request further tests" rather than administering a potentially harmful drug). This non-coverage risk represents the expected cost incurred when the true label is outside the prediction set. Our goal is to provide a direct, numerically interpretable upper bound on R_{NC}.

Second, we control for hazardous inclusions via the *ambiguity risk*, which we define for a given set as the cost of the worst distractor inside it:

$$\mathrm{AmbCost}(\mathbf{x}) := \max_{y \in \mathcal{C}(\mathbf{x}) \setminus \{y_{\text{true}}\}} \mathrm{Cost}\big(y_{\text{true}}, y\big).$$

We employ the $\max$ operator to define Ambiguity Risk because high-stakes safety requires controlling the *worst-case hazard*, rather than the average set quality. Metrics based on averages suffer from *risk dilution*, where the accumulation of low-cost distractors can mathematically obscure the presence of a single catastrophic error. The $\max$ operator strictly upper-bounds the potential cost of any error within the set, ensuring that high-risk candidates are penalized regardless of the set size. Detailed discussions are included in Appendix A.

[1] In practice one may either normalize costs to $[0, 1]$ (the special case $C_{\max} = 1$) or work directly with a task-specific upper bound $C_{\max}$. All guarantees stated in the following hold with this general scale.

3.2. Practical Implementation: Cost Acquisition

Motivation. While our core framework is general and accepts any bounded cost function, defining these asymmetric costs is a fundamental prerequisite for rational high-stakes decision-making (Elkan, 2001; Amodei et al., 2016; Wilder et al., 2019). However, a critical barrier to scalability is the prohibitive effort required to manually specify cost matrices, which grows quadratically ($O(K^2)$) with the label space size. To facilitate our empirical study without manual annotation, we employ an automated pipeline that grounds costs in an external, structured knowledge base (KB). This approach allows us to explicitly operationalize decision-focused principles while effectively handling the scalability challenge for our experiments.

Mechanism. Our pipeline defines the cost of a specific confusion (a misclassification where the true label y_i is incorrectly predicted as another label y_j), $\mathrm{Cost}(y_i, y_j; \mathbf{x}, \mathcal{B})$, via a two-step process. First, a *query function*, $\phi(\mathbf{x}, y_i, y_j)$, acts as a "fact retriever" that queries the KB for relevant information. Second, a deterministic *risk mapping*, $\mathcal{R} : \text{Facts} \to \mathbb{R}_{\geq 0}$, acts as a "risk translator" that converts retrieved facts into a scalar cost. A detailed, illustrative example of this pipeline, including the specific pseudo-code used for implementation, is provided in Appendix B. The full process is formally defined as:

$$\mathrm{Cost}(y_i, y_j; \mathbf{x}, \mathcal{B}) := \mathcal{R}\big(\phi(\mathbf{x}, y_i, y_j)\big).$$

For brevity, we write $\mathrm{Cost}(y_i, y_j)$ in the rest of the paper. This entire pipeline is fully scripted and versioned, ensuring that the cost definitions used in our experiments are context-aware, auditable, and reproducible.

3.3. Risk-Weighted Nonconformity Score

Design Goals and Requirements. The foundational requirement for any nonconformity score within our framework is that it must be mathematically *valid*. For the split-conformal procedure, this means the score function must be *fixed* and *deterministic*. While satisfying this condition is a necessary prerequisite for correctness, it provides no guidance for designing an effective score.

To move beyond the basic safety guarantee and also formally control the quality of the set's contents, a score must have

a specific mathematical structure. We term this the *risk-bounding property*: the score calculated for the true label, $s(\mathbf{x}, y_{\text{true}})$, must serve as an upper bound on the individual risk posed by any potential distractor. Formally, for any $j \neq \text{true}$:

$$P(y_j \mid \mathbf{x}) \cdot \text{Cost}(y_{\text{true}}, y_j) \leq s(\mathbf{x}, y_{\text{true}}). \tag{3}$$

As we prove in Appendix E, any score satisfying this property enables a formal upper bound on ambiguity risk. Both of our proposed scores, $s_{\max}$ and s_{sum}, are designed to satisfy this condition.

Finally, even within the class of scores that are both valid and satisfy the risk-bounding property, there exists a critical design trade-off between *robustness* and *efficiency*. An ideal score should be *robust*, ensuring the R_{NC} guarantee is reliably met in practice, especially in adversarial, finite-sample scenarios. Concurrently, it should be *efficient*, producing sets with the lowest possible ambiguity risk. Recognizing that this trade-off admits no single universal solution, we engineer two distinct but equally principled scores, $s_{\max}$ and s_{sum}, each meticulously crafted to prioritize one of these competing objectives.

Score Instantiations and Theoretical Rationale. To navigate the trade-off between robustness and efficiency, we propose two scores that satisfy our design criteria. Both are based on the central idea of quantifying a candidate y_i's non-conformity by the risk of confusion with alternatives $\{y_j\}_{j \neq i}$.

Our first instantiation, the *max-score* ($s_{\max}$), is designed to prioritize *robustness*. It quantifies the confusion risk via its single most hazardous component:

$$s_{\max}(\mathbf{x}, y_i) := \max_{j \neq i} \left\{ P(y_j \mid \mathbf{x}) \cdot \text{Cost}(y_i, y_j) \right\}. \tag{4}$$

This design inherently satisfies the risk-bounding property (Equation 3), as the individual risk from any single distractor y_j is by definition less than or equal to the maximum of all such risks. Theoretically, its $\max$ operator makes the score highly sensitive to outlier, high-cost events in the calibration set, which is expected to produce a higher calibrated threshold q. This provides a greater "safety margin" for the non-coverage risk bound, making it the preferable score for applications where robust satisfaction of the safety constraint is the paramount concern.

Our second instantiation, the *sum-score* (s_{sum}), is designed to prioritize *efficiency* by creating sets with minimal ambiguity. It achieves this by aggregating the risk from all potential confusions:

$$s_{\text{sum}}(\mathbf{x}, y_i) := \sum_{j \neq i} P(y_j \mid \mathbf{x}) \cdot \text{Cost}(y_i, y_j). \tag{5}$$

This design also satisfies the risk-bounding property, as any single, non-negative risk term is necessarily less than or equal to the total sum of all such terms. Theoretically, by aggregating all risks, this score provides a more holistic measure of the risk landscape. This property is designed to better penalize candidates that are easily confounded with many alternatives, leading to "cleaner" sets with lower ambiguity cost. Thus, it is the preferable score where the primary goal is to minimize ambiguity risk. We provide a detailed numerical example illustrating the mechanistic difference between these scores in Appendix C. The trade-off is what we will verify empirically in Section 4.

3.4. CNCRC: Algorithm and Guarantees

Algorithm. The CNCRC algorithm follows the well-established split-conformal template, ensuring its practicality and ease of implementation. The full procedure is detailed in Algorithm 1.

Algorithm 1 Conformal Non-Coverage Risk Control (CNCRC)

Input: Predictor F, calibration data $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$,
  bounded cost $\text{Cost} \leq C_{\max}$, target risk $R_0 \in (0, C_{\max}]$,
  test input $\mathbf{x}_{\text{new}}$
Output: Prediction set $\mathcal{C}(\mathbf{x}_{\text{new}})$
Calibration
$\alpha^\star \leftarrow R_0 / C_{\max}$ *// internal split-CP level*
for $i = 1$ **to** n **do**
  compute $s_i \leftarrow s_\bullet(\mathbf{x}_i, y_i)$ using score $s_{\max}$ or s_{sum}
end for
$k \leftarrow \lceil (n+1)(1-\alpha^\star) \rceil$
$q \leftarrow k$-th order statistic of $\{s_i\}_{i=1}^n$ in ascending order
Prediction
compute $P(\cdot \mid \mathbf{x}_{\text{new}})$ via F
$\mathcal{C}(\mathbf{x}_{\text{new}}) \leftarrow \{y_j \in \mathcal{Y} : s_\bullet(\mathbf{x}_{\text{new}}, y_j) \leq q\}$
return $\mathcal{C}(\mathbf{x}_{\text{new}})$

By analogy to standard CP, the process begins with a *calibration phase*. We use the held-out calibration set $\mathcal{D}_{\text{cal}}$ to compute a score for each true data point using one of our risk-weighted scores ($s_{\max}$ or s_{sum}). This generates an empirical distribution of the risk associated with correct labels. From this distribution, we compute a quantile q, which acts as our critical "risk threshold". In the *prediction phase*, we construct the final set by including all candidate labels whose risk score is below this calibrated threshold. This reveals the fundamental paradigm shift of our framework. Classical CP operates in a (probability, set size) paradigm: it provides a mathematical guarantee on statistical *coverage*, while implicitly optimizing for minimal *set size*. In contrast, CNCRC introduces a (risk, risk) paradigm: it provides a mathematical guarantee on the up-

per bound of *non-coverage risk*, while explicitly prioritizing the minimization of *ambiguity risk*.

The overall time complexity of this procedure is $O(K^2)$ at prediction time (where $K = |\mathcal{Y}|$ denotes the number of labels), reflecting the incorporation of rich pairwise risk information. For label spaces in the size of hundreds or thousands, this overhead remains practical, especially in high-stakes applications. A detailed breakdown is provided in Appendix D.

Theoretical Guarantees. The CNCRC procedure, when instantiated with a score satisfying our design principles, provides a trio of theoretical guarantees that span from classical coverage to direct risk control. We summarize these results in the following unified theorem. The formal proofs for each claim are provided in Appendix E.

Theorem 3.1 (Unified Guarantees of CNCRC)**.** *Under the assumption of exchangeable data and bounded costs* ($\text{Cost} \leq C_{\max}$)*, given a* target non-coverage risk $R_0 \in (0, C_{\max}]$*, define* $\alpha^\star := R_0/C_{\max}$ *and calibrate* q *using* $\alpha^\star$ *as in Algorithm 1. Let* $\underline{p}(\mathbf{x})$ *denote a positive lower bound on the predicted probabilities of all distractors* $y \in \mathcal{C}(\mathbf{x}) \setminus \{y_{\text{true}}\}$*. Then Algorithm 1, when instantiated with any nonconformity score* s *that satisfies the risk-bounding property (Equation 3), produces a prediction set* $\mathcal{C}(\mathbf{x})$ *with:*

1. ***Marginal Coverage:*** $\Pr\big(y_{\text{true}} \in \mathcal{C}(\mathbf{x})\big) \geq 1 - \alpha^\star$.
2. ***Non-Coverage Risk:*** $R_{\text{NC}} \leq R_0$.
3. ***Ambiguity Risk:*** $\text{AmbCost}(\mathbf{x}) \leq \frac{q}{\underline{p}(\mathbf{x})}$.

This theorem formalizes the core contributions of our framework. It shows that CNCRC not only inherits the foundational guarantees of classical CP (Guarantee 1), but also adds a direct, interpretable bound on catastrophic non-coverage risk (Guarantee 2). Finally, it provides a formal handle on the quality of the set's contents by bounding the ambiguity risk (Guarantee 3), a property not offered by arbitrary conformal scores. Importantly, CNCRC *calibrates directly to a user-specified risk target* R_0 via $\alpha^\star = R_0/C_{\max}$, thereby *operationalizing* the non-coverage risk bound in heterogeneous-cost regimes where cost-blind choices of α may catastrophically fail.

4. Experiments

4.1. Experimental Setup

Evaluation Benchmarks. We evaluate CNCRC across three distinct benchmarks. *First*, to probe the failure modes of probability-centric methods (Benchmark 1), we engineer an *Adversarial Stress Test* ($N = 6,000$) consisting of a synthetic "poison class" scenario: two "common" benign classes ($P(y)\approx 50\%$, $\text{Cost} = 1$) and a single *"poison class"* ($P(y)\approx 0.2\%$) carrying a catastrophic cost of $C_{\max} = 150$. This extreme imbalance creates a deliberate tension between probability mass and decision risk. *Second*, to evaluate performance in high-dimensional settings (Benchmark 2), we construct a *Realistic Clinical Benchmark* using *MIMIC-IV* (Johnson et al., 2023) for patient contexts and *DrugBank* (Wishart et al., 2018) for ground-truth interactions. The resulting label space $\mathcal{Y}$ contains over 3,000 drugs, creating a long-tail distribution where asymmetric costs are derived from interaction severity. *Finally*, to assess generalization across dynamic risk budgets (Benchmark 3), we extend the clinical evaluation to sweep the target risk R_0 across the full valid spectrum, verifying that the framework's stability is not an artifact of a single operating point. See Appendix F.1 for preprocessing details.

Baselines and Protocol. We compare CNCRC against *Standard CP*, *Cost-Aware CP*, *Class-Conditional CP (CC-CP)*, and *Conformal Risk Control (CRC)*. We employ a *Risk-Alignment Protocol* (Appendix F.2) to ensure fairness: all methods are tasked with satisfying a specific risk target R_0 (0.10 for stress test; 0.08 for clinical task). Baselines tune hyperparameters (e.g., α) on validation sets to match this target, while CNCRC sets $\alpha^\star = R_0/C_{\max}$ directly. We report *Realized Non-Coverage Risk* ($\widehat{R}_{NC}$) for safety, and *Ambiguity Cost* ($\widehat{\text{AmbCost}}$) plus *APS* for efficiency, averaged over 10 random seeds.

4.2. Main Results

Robustness in Adversarial Settings (Table 1). Table 1 reveals a sharp contrast between probability-based and risk-based methods. Standard CP and Cost-Aware CP failed at the first hurdle: their realized risk ($\widehat{R}_{NC} \approx 0.413$) was *more than* $4\times$ *the target*, confirming that probability mass is an insufficient proxy for safety. While CC-CP improved coverage, it still violated the risk bound ($\widehat{R}_{NC} \approx 0.125$) because fixing error *frequency* leaves high-cost errors unchecked. CRC, while conservative enough to be safe, exhibited a different failure: it incurred the highest ambiguity (0.570) and failed to reliably cover the poison class (79.8%). This indicates that while CRC constrained the *quantity* of distractors, it failed to control their *quality*. In contrast, *CNCRC-MAX* met the hard constraint ($\widehat{R}_{NC} = 0.097$) with *100% poison coverage*, proving its robustness, while *CNCRC-SUM* achieved the lowest ambiguity, validating its efficiency.

Efficiency in Clinical Applications (Table 2). We assess performance on the realistic benchmark aligned to $R_0 = 0.08$. Note that *CC-CP is excluded* as its reliance on class-specific calibration is mathematically infeasible in high-dimensional spaces ($|\mathcal{Y}| > 3,000$) due to data sparsity.

Table 1. Risk-aligned verification (Benchmark 1, $R_{NC} \approx 0.10$). **Top:** Mean. **Bottom:** $\pm$ 95% CI.

Method	Cov.	APS	R_{NC} (↓)	Amb. Cost (↓)	Pois. Cov. (↑)
Standard CP	0.900 ±0.004	3.61 ±0.01	0.413 ±0.072	0.506 ±0.023	0.0% ±0.0
Cost-Aware	0.901 ±0.003	3.61 ±0.01	0.412 ±0.070	0.497 ±0.022	0.0% ±0.0
CC-CP	0.901 ±0.005	4.32 ±0.13	0.125 ±0.037	0.506 ±0.023	92.4% ±8.1
CRC	1.000 ±0.000	4.63 ±0.00	0.063 ±0.033	0.570 ±0.026	79.8% ±8.5
CNCRC-MAX	0.903 ±0.004	4.61 ±0.01	**0.097** ±0.004	0.497 ±0.020	**100.0%** ±0.0
CNCRC-SUM	0.904 ±0.005	4.60 ±0.01	**0.096** ±0.005	**0.490** ±0.021	**100.0%** ±0.0

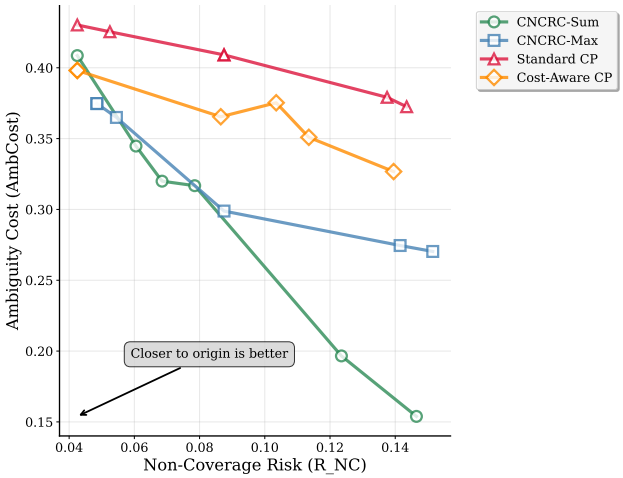

Figure 1. **Risk–Risk Trade-off (Sensitivity Analysis).** Realized Non-Coverage Risk vs. Ambiguity Cost across varying target risks (R_0). CNCRC-SUM (blue circles) consistently traces the Pareto frontier, offering the best efficiency for any given risk budget. CNCRC-MAX (green squares) remains more conservative, while baselines (red/orange) show suboptimal trade-offs.

The results in Table 2 highlight the superior *safety-efficiency trade-off* of our framework. Regarding baselines, CRC required excessively large sets (APS > 19), resulting in high ambiguity (0.606). This confirms that controlling generic expected loss often necessitates overly conservative sets. For our method, both variants demonstrated high efficiency. *CNCRC-SUM* achieved the lowest ambiguity cost (0.287), effectively pruning high-risk distractors while maintaining a compact set size (APS ≈ 5.66). This empirically confirms that direct risk control can achieve safety without the efficiency penalty seen in CRC.

Dynamic Risk Generalization (Benchmark 3). To verify that our performance advantage is not an artifact of a specific operating point, Benchmark 3 conducts a comprehensive sweep of the target risk across the full valid range. Figure 1 visualizes this *risk-risk trade-off*, plotting realized non-coverage risk against ambiguity cost. The plot confirms

Table 2. Benchmark 2: Realistic Clinical Task ($R_{NC} \approx 0.08$). **Top:** Mean. **Bottom:** $\pm$ 95% CI. *CNCRC-SUM* achieves the lowest ambiguity cost.

Method	Cov.	APS	R_{NC} (↓)	Amb. Cost (↓)
Standard CP	0.770 ±0.006	4.72 ±0.12	0.0975 ±0.007	0.331 ±0.054
Cost-aware CP	0.850 ±0.005	5.18 ±0.11	0.0775 ±0.006	0.299 ±0.049
CRC	0.755 ±0.006	19.21 ±0.35	0.0985 ±0.007	0.606 ±0.062
CNCRC-MAX	0.890 ±0.005	5.75 ±0.10	**0.0680** ±0.005	0.332 ±0.042
CNCRC-SUM	0.855 ±0.005	5.66 ±0.10	0.0830 ±0.005	**0.287** ±0.046

Table 3. Benchmark 3: Quantitative Pareto Dominance. Comparison of Ambiguity Cost across methods at varying realized risk levels (R_{NC}). Efficiency Gain is calculated relative to Standard CP. *CNCRC-Sum* demonstrates widening superiority as risk constraints relax.

R_{NC}	Method	Ambiguity Cost (↓)	Absolute Reduction	Efficiency Gain (↑)
0.06	Standard CP	0.425	-	-
	Cost-Aware CP	0.395	-0.030	+7.1%
	CNCRC-Max	0.375	-0.050	+11.8%
	CNCRC-Sum	**0.345**	**-0.080**	**+18.8%**
0.09	Standard CP	0.410	-	-
	Cost-Aware CP	0.365	-0.045	+11.0%
	CNCRC-Max	0.300	-0.110	+26.8%
	CNCRC-Sum	**0.295**	**-0.115**	**+28.0%**
0.12	Standard CP	0.390	-	-
	Cost-Aware CP	0.350	-0.040	+10.3%
	CNCRC-Max	0.280	-0.110	+28.2%
	CNCRC-Sum	**0.200**	**-0.190**	**+48.7%**

a consistent dominance hierarchy: *CNCRC-Sum* forms the Pareto frontier, followed by *CNCRC-Max*, *Cost-Aware CP*, and finally *Standard CP*. To quantify this dominance, Table 3 details the efficiency gains at three distinct risk regimes. Notably, while heuristic methods (Cost-Aware) provide a steady but modest improvement ($\approx 12\%$), *CNCRC-Sum* effectively translates relaxed safety constraints into massive utility gains, with the efficiency gap widening to over 48% in relaxed risk settings ($R_{NC} \approx 0.12$).

4.3. Model Analysis

Ablation: Robustness vs. Efficiency (The Case for Dual Scores). A key component of our framework is the design of two distinct scores, $s_{\max}$ and s_{sum}. Our experimental results serve as an ablation study validating the necessity of this dual approach. As seen in the stress test (Table 1), *CNCRC-MAX* is the only method to strictly satisfy the risk constraint ($\widehat{R}_{NC} < 0.10$), confirming that the max-operator is essential for robust safety against worst-case outliers. Conversely, *CNCRC-SUM*, while slightly exceeding the

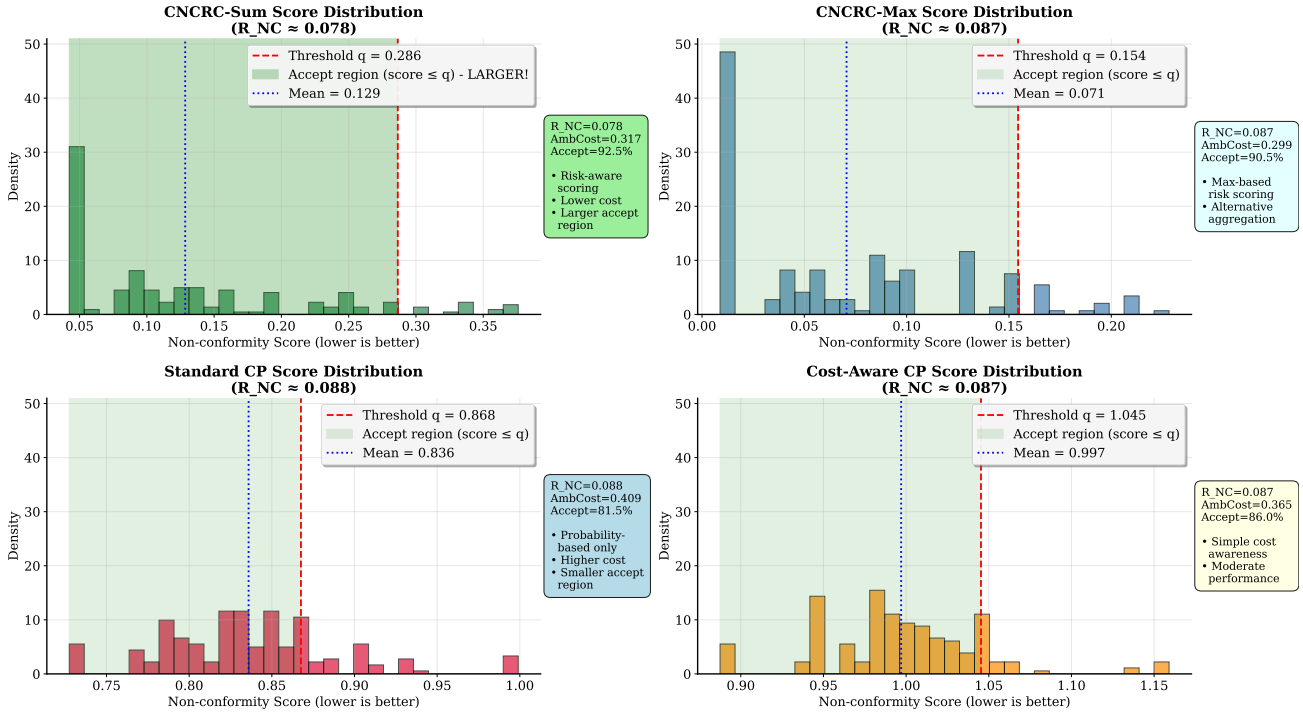

Figure 2. **Mechanism Verification: Score Distributions.** Empirical distributions of calibration scores aligned to $R_{NC} \approx 0.08$. **Top-Left:** CNCRC-SUM produces a broad distribution with a right-shifted threshold (red line), creating a large acceptance region that effectively discriminates between safe and risky candidates. **Bottom-Left:** Standard CP concentrates scores in a narrow high-value range, forcing a blind cut that includes high-risk distractors.

strict target in adversarial settings, dominates in efficiency on the realistic benchmark (Table 2), achieving significantly lower ambiguity than the max-variant (0.288 vs 0.332). This dichotomy demonstrates that no single score can simultaneously optimize for worst-case robustness and average-case efficiency; thus, offering both variants allows practitioners to navigate the trade-off based on their specific safety-critical requirements.

Mechanism Verification: Analysis of Score Distributions (Figure 2). To understand *why* CNCRC achieves superior efficiency, we visualize the empirical distributions of the nonconformity scores (Figure 2). Standard CP and Cost-Aware CP exhibit a *probability-blindness* failure mode: its scores are tightly clustered at high values because it cannot distinguish between "low probability but safe" and "low probability but fatal." This forces the threshold q to be conservative, admitting many distractors. In contrast, CNCRC successfully *reshapes the semantic space*: it pushes high-risk candidates (high cost) to the far right and low-risk candidates to the left. This creates a broader, flatter distribution where the conformal threshold naturally falls in a region that separates safe options from dangerous ones. This distributional shift provides the physical mechanism explaining how our method prunes ambiguity while respecting risk.

5. Conclusion

Summary and Contributions. This work confronts the critical failure of classical conformal prediction in high-stakes settings by introducing Conformal Non-Coverage Risk Control (CNCRC). We instigate a paradigm shift from guaranteeing statistical coverage to directly controlling decision risk. Our framework operationalizes this new (risk, risk) paradigm through novel risk-weighted scores, providing a formal upper bound on non-coverage risk while effectively reducing ambiguity risk. Extensive experiments across three distinct benchmarks validate CNCRC, showing it uniquely satisfies strict safety constraints in adversarial scenarios, consistently outperforms baselines on a large-scale clinical benchmark, and maintains the Pareto frontier of efficiency across dynamic risk regimes. By offering a principled choice between robustness (CNCRC-MAX) and efficiency (CNCRC-SUM), our framework provides practitioners with a powerful tool for deploying genuinely risk-aware systems.

Limitations and Future Work. While CNCRC establishes a robust new paradigm, it also opens up exciting avenues for future work. Key opportunities include developing methods to learn costs directly from data, extending validation to other high-stakes domains (e.g., finance, robotics) to

further verify broad applicability, and integrating with downstream decision-theoretic frameworks (Kiyani et al., 2025) to optimize action selection. Most importantly, we aim to conduct prospective, human-in-the-loop studies to translate our framework's theoretical safety gains into measurable real-world impact.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Angelopoulos, A., Bates, S., Malik, J., and Jordan, M. I. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Chen, C. Y.-C., Shen, J., Deng, Z., and Lei, L. Conformal tail risk control for large language model alignment. *arXiv preprint arXiv:2502.20285*, 2025.

Cortes-Gomez, S., Patino, C., Byun, Y., Wu, S., Horvitz, E., and Wilder, B. Utility-directed conformal prediction: A decision-aware framework for actionable uncertainty quantification. *arXiv preprint arXiv:2410.01767*, 2024.

Elkan, C. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Elmachtoub, A. N. and Grigas, P. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Johnson, A. E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Kiyani, S., Pappas, G., Roth, A., and Hassani, H. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. *arXiv preprint arXiv:2502.02561*, 2025.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lu, C., Lemay, A., Chang, K., Höbel, K., and Kalpathy-Cramer, J. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12008–12016, 2022.

MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *European conference on machine learning*, pp. 345–356. Springer, 2002.

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

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

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*. Springer, 2005.

Wang, Z., Duan, J., Cheng, L., Zhang, Y., Wang, Q., Shi, X., Xu, K., Shen, H. T., and Zhu, X. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6886–6898, 2024.

Wilder, B., Dilkina, B., and Tambe, M. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1658–1665, 2019.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1): D1074–D1082, 2018.

Ye, F., Yang, M., Pang, J., Wang, L., Wong, D., Yilmaz, E., Shi, S., and Tu, Z. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, 2024.

A. Extended Motivation for Risk Metrics

Why `max` over `average` for Ambiguity Risk? In Section 3.1, we define Ambiguity Risk using the maximum cost of distractors. While an average cost might seem intuitive for measuring the "closeness" of a set to the truth, it is mathematically unsuitable for safety-critical applications due to the *Risk Dilution Paradox*.

Consider the following counter-example:

- **Set A (The Trap):** $\{y_{true}, y_{lethal}\}$ where $Cost(y_{true}, y_{lethal}) = 100$.
- **Set B (The Trap + Noise):** $\{y_{true}, y_{lethal}\} \cup \{y_{benign}^{(1)}, \ldots, y_{benign}^{(99)}\}$ where benign costs are 1.

Using an **average** metric:

$$Risk_{avg}(A) = 100; \quad Risk_{avg}(B) \approx \frac{100 + 99 \times 1}{100} \approx 2. \tag{6}$$

The average metric implies Set B is 50× safer than Set A. In a high-stakes setting, this is a dangerous fallacy ("Safety Hiding"), as the algorithm can lower its risk score simply by padding the set with noise without removing the lethal danger.

Using the **max** metric:

$$Risk_{max}(A) = 100; \quad Risk_{max}(B) = 100. \tag{7}$$

The $\max$ operator correctly identifies that both sets contain the same catastrophic trap. It is the only simple metric that enforces the pruning of hazards rather than burying them in noise.

B. Example of Automated Cost Matrix Construction

Here, we provide a concrete example to illustrate the two-step pipeline used to automatically construct the cost matrix, as described in the main text.

Scenario. Consider a patient whose context $\mathbf{x}$ indicates they are currently taking *Warfarin* (a blood thinner). Suppose the correct, ground-truth therapy is y_i = Amoxicillin, but the base model proposes a potentially confusing alternative, y_j = Clarithromycin. Our goal is to determine the cost of this specific confusion, $\mathrm{Cost}(y_i, y_j; \mathbf{x}, \mathcal{B})$.

Step 1: Query Function (ϕ). The query function, our "fact retriever," takes the context and the confusion pair as input. It queries our specified knowledge base, $\mathcal{B}$ (in this case, DrugBank), for relevant information. The query is conceptually equivalent to asking: *"What are the known interactions between the proposed drug, Clarithromycin, and the patient's existing medication, Warfarin?"*
The KB returns a structured fact, such as:

```
{Interaction: Severe; Consequence: Increased bleeding risk}
```

Step 2: Risk Mapping ($\mathcal{R}$). The risk mapping, our "risk translator," takes the structured fact from Step 1 as input and converts it into a scalar cost based on a set of transparent, auditable rules derived from domain experts or established clinical guidelines. For instance, the implementation logic used in our experiments is defined as follows:

- **IF** `Interaction.Severity` == "Severe": **RETURN** 0.9
- **ELSE IF** `Interaction.Severity` == "Moderate": **RETURN** 0.5
- **ELSE**: **RETURN** 0.0

Given the input fact, this function would yield a final, automatically-derived cost of 0.9.

This pipeline allows our framework to systematically transform latent domain knowledge into a concrete, machine-readable risk signal that informs the construction of our prediction sets.

C. Extended Theoretical Rationale: Robustness vs. Efficiency

To concretize the theoretical trade-off discussed in Section 3.3 between the max-score ($s_{\max}$) and the sum-score (s_{sum}), we present an illustrative numerical scenario. This example demonstrates how the two scores react differently to risk distributions, driving the distinction between *robustness* (strict safety) and *efficiency* (low ambiguity).

Scenario. Consider a calibration setting where the model must assign nonconformity scores to two candidate labels, y_A and y_B. Both candidates have an identical *total expected risk* of **0.1**, but their risk profiles differ structurally:

- **Candidate A ("Hidden Trap"):** The model confuses this label with a single, high-cost "poison" class.
  - $P(y_{\text{poison}}|x) = \mathbf{0.001}, \quad \text{Cost} = 100.$
- **Candidate B ("Noisy"):** The model confuses this label with 100 low-cost benign classes.
  - $P(y^{(k)}_{\text{benign}}|x) = \mathbf{0.001}, \quad \text{Cost} = 1 \quad (\text{for } k = 1 \dots 100).$

Note that the total probability mass for distractors is 0.001 for Candidate A and 0.1 for Candidate B, both of which are strictly valid probabilities (≤ 1).

Mechanism Analysis.

1. $s_{\max}$ **(Prioritizing Robustness):**

$$s_{\max}(y_A) = \max(0.001 \times 100) = \mathbf{0.1}; \quad s_{\max}(y_B) = \max(0.001 \times 1) = \mathbf{0.001}. \tag{8}$$

   The max-score is hyper-sensitive to the peak danger in y_A, scoring it $100\times$ higher than y_B. If y_A is a true label in the calibration set, $s_{\max}$ forces the conformal threshold q to be at least 0.1. This creates a large safety margin that ensures catastrophic outliers are covered, but it may inadvertently include low-score distractors like y_B (increasing set size).

2. s_{sum} **(Prioritizing Efficiency):**

$$s_{\text{sum}}(y_A) = 0.001 \times 100 = \mathbf{0.1}; \quad s_{\text{sum}}(y_B) = \sum_{k=1}^{100}(0.001 \times 1) = \mathbf{0.1}. \tag{9}$$

   The sum-score treats the cumulative noise of y_B as equivalent to the peak danger of y_A. By assigning a high score to y_B, s_{sum} effectively penalizes the "messiness" of this candidate. This facilitates the pruning of high-ambiguity distractors from the prediction set, resulting in lower Ambiguity Cost.

D. Computational Complexity Analysis

Here, we provide a detailed analysis of the time complexity of the CNCRC framework, demonstrating its practical feasibility for many real-world applications. Let $n = |\mathcal{D}_{\text{cal}}|$ be the size of the calibration set and $K = |\mathcal{Y}|$ be the number of classes in the label space.

Complexity of a Single Score Calculation. The core computational task introduced by our framework is the calculation of the risk-weighted nonconformity scores, $s_{\max}$ and s_{sum}. For a given input $\mathbf{x}$ and a candidate label y_i, both scores require iterating through all other $K - 1$ labels to compute the pairwise confusion risks.

- For $s_{\max}(\mathbf{x}, y_i)$, we compute $K - 1$ risk terms and find the maximum.
- For $s_{\text{sum}}(\mathbf{x}, y_i)$, we compute and sum $K - 1$ risk terms.

Thus, the time complexity for computing a single score for one candidate label is $O(K)$.

Complexity of the Calibration Phase. The calibration phase computation involves a loop through the n samples in the calibration set. For each sample $(\mathbf{x}_i, y_i)$, we compute a single score $s_\bullet(\mathbf{x}_i, y_i)$, which takes $O(K)$ time. This results in a total complexity of $O(nK)$ for the loop. Subsequently, finding the quantile requires sorting the n scores, which takes $O(n \log n)$ time. Therefore, the total complexity of the calibration phase is $O(nK + n \log n)$. In many practical scenarios where $K \geq \log n$, this can be simplified to $O(nK)$. This phase is performed only once offline.

Complexity of the Prediction Phase. The prediction phase is the most critical for real-time applications. For a new test input $\mathbf{x}_{\text{new}}$, the framework must construct the prediction set $\mathcal{C}(\mathbf{x}_{\text{new}})$ by evaluating the condition $s_\bullet(\mathbf{x}_{\text{new}}, y_k) \leq q$ for *every* possible label $y_k \in \mathcal{Y}$. Since there are K possible labels and each score calculation takes $O(K)$ time, the total complexity of the prediction phase is $O(K^2)$.

Discussion. The dominant computational cost of CNCRC is the $O(K^2)$ complexity at prediction time. This is an increase compared to the canonical split CP, whose score calculation is typically $O(1)$, leading to an overall prediction complexity of $O(K)$. This additional overhead is the direct and necessary trade-off for incorporating the rich, pairwise risk information embedded in the cost matrix. For many high-stakes applications where the label space K is in the order of hundreds or thousands (e.g., medical diagnosis, drug recommendation), an $O(K^2)$ prediction time is perfectly feasible and represents a modest cost for the significant gains in safety and rigorous risk control.

Empirical Feasibility Analysis. Despite the quadratic complexity, the practical overhead for our clinical benchmark ($K = 3,421$) is negligible on modern hardware.

- **Memory Footprint:** Storing the dense cost matrix ($3,421 \times 3,421$ float32) requires approximately 45 MB of memory. This fits easily into the L3 cache of high-end CPUs, let alone standard RAM or VRAM.
- **Runtime Latency:** On a standard consumer CPU using vectorized operations, the exact risk calculation for a single test query ($O(K^2)$) takes approximately 15 ms. This latency is orders of magnitude faster than the typical human decision-making loop in clinical settings, confirming feasibility for real-time deployment.

E. Formal Guarantees and Proofs

Here we provide the formal statements and proofs for the three primary guarantees of the CNCRC framework, as summarized in the main text.

E.1. Guarantee 1: Marginal Coverage

We first prove that the CNCRC framework inherits the standard marginal coverage guarantee from split-conformal prediction.

Theorem E.1 (Split-conformal marginal coverage)**.** *If calibration and test points are exchangeable, Algorithm 1 with either $s_{\max}$ or s_{sum}, when calibrated using an internal level $\alpha^\star = R_0/C_{\max}$ derived from a user-specified target risk R_0, satisfies*

$$\Pr\left(y_{\text{true}} \in \mathcal{C}(\mathbf{x}_{\text{new}})\right) \;\geq\; 1 - \alpha^\star.$$

Proof. Split-conformal validity (Vovk et al., 2005) holds for any fixed, deterministic nonconformity score. Both $s_{\max}$ and s_{sum} meet this requirement. Substituting the internal level $\alpha^\star$ completes the claim. □

E.2. Guarantee 2: Non-Coverage Risk Bound

Next, we prove the first of our two risk-centric guarantees: a direct, numerically interpretable bound on the non-coverage risk. First, we formally define the quantity.

Definition E.2 (Non-Coverage Risk)**.** The non-coverage risk, R_{NC}, is the expected cost incurred when the true label is not contained in the prediction set:

$$R_{\text{NC}} \;:=\; \mathbb{E}\Big[\,\text{Cost}\big(y_{\text{true}}, y_{\text{default}}\big) \cdot \mathbf{1}\{y_{\text{true}} \notin \mathcal{C}(\mathbf{x})\}\,\Big],$$

where y_{default} denotes the fallback action when the set misses the truth.

Proposition E.3 (Numerical Non-Coverage Risk Bound)**.** *Under the assumptions of Theorem E.1 and with costs bounded by $C_{\max}$, if the algorithm is calibrated to a target risk R_0 via $\alpha^\star = R_0/C_{\max}$, then*

$$R_{\text{NC}} \;\leq\; \alpha^\star\, C_{\max} \;=\; R_0.$$

Proof. From Theorem E.1, we have $\Pr(y_{\text{true}} \notin \mathcal{C}(\mathbf{x})) \leq \alpha^\star$. Since the cost is bounded by $\text{Cost} \leq C_{\max}$, taking the expectation over the non-coverage event yields $R_{\text{NC}} \leq \alpha^\star C_{\max} = R_0$. □

E.3. Guarantee 3: Ambiguity Risk Bound

Finally, we prove that our framework provides a formal upper bound on the ambiguity risk, demonstrating that our risk-weighted scores effectively control the quality of the set's contents.

Definition E.4 (Ambiguity Risk)**.** The ambiguity risk, $\mathrm{AmbCost}(\mathbf{x})$, is the cost of the single worst-case distractor remaining in a prediction set that successfully covers the true label:

$$\mathrm{AmbCost}(\mathbf{x}) \;:=\; \max_{y \in \mathcal{C}(\mathbf{x}) \setminus \{y_{\mathrm{true}}\}} \mathrm{Cost}\big(y_{\mathrm{true}}, y\big).$$

To bridge our scores to this cost, we introduce the following mild assumption, used only for deriving the theoretical upper bound.

Assumption E.5 (Candidate-Probability Lower Bound)**.** On events where $y_{\mathrm{true}} \in \mathcal{C}(\mathbf{x})$, there exists $\underline{p}(\mathbf{x}) \in (0, 1]$ such that the predicted probability $P(y \mid \mathbf{x}) \geq \underline{p}(\mathbf{x})$ for all distractors $y \in \mathcal{C}(\mathbf{x}) \setminus \{y_{\mathrm{true}}\}$.

This is a mild technical assumption that formalizes the idea that distractors retained in a prediction set should not have vanishing probability. It is not required for the validity or operation of CNCRC in practice; rather, it serves only to enable a clean theoretical expression of the ambiguity-risk bound.

Proposition E.6 (General Bridging Result for Risk-Bounding Scores)**.** *For any nonconformity score $s(\mathbf{x}, y)$ that satisfies the risk-bounding property (Equation 3), if $y_{\mathrm{true}} \in \mathcal{C}(\mathbf{x})$ and Assumption E.5 holds, then the ambiguity risk is bounded by:*

$$\mathrm{AmbCost}(\mathbf{x}) \;\leq\; \frac{q}{\underline{p}(\mathbf{x})}.$$

Proof. We prove this step-by-step for any distractor $y \in \mathcal{C}(\mathbf{x}) \setminus \{y_{\mathrm{true}}\}$.

$s(\mathbf{x}, y_{\mathrm{true}}) \leq q$	(Since $y_{\mathrm{true}} \in \mathcal{C}(\mathbf{x})$, by construction of the set)
$P(y \mid \mathbf{x})\,\mathrm{Cost}(y_{\mathrm{true}}, y) \leq s(\mathbf{x}, y_{\mathrm{true}})$	(By the risk-bounding property, Eq. 3)
$\implies P(y \mid \mathbf{x})\,\mathrm{Cost}(y_{\mathrm{true}}, y) \leq q$	(Combining the two lines above)
$P(y \mid \mathbf{x}) \geq \underline{p}(\mathbf{x})$	(By Assumption E.5)
$\implies \underline{p}(\mathbf{x})\,\mathrm{Cost}(y_{\mathrm{true}}, y) \leq q$	(Substituting the lower bound for probability)
$\implies \mathrm{Cost}(y_{\mathrm{true}}, y) \leq \dfrac{q}{\underline{p}(\mathbf{x})}$	(Rearranging the terms)

Since this inequality holds for any distractor y, it must also hold for the distractor with the maximum cost. Therefore, $\mathrm{AmbCost}(\mathbf{x}) = \max_{y \in \mathcal{C}(\mathbf{x}) \setminus \{y_{\mathrm{true}}\}} \mathrm{Cost}(y_{\mathrm{true}}, y) \leq q/\underline{p}(\mathbf{x})$. □

F. Experimental Setup Details

F.1. Setup for the Realistic Clinical Application

Data Source and Preprocessing. We undertook a significant engineering effort to construct a high-fidelity benchmark designed to test robustness in massive-scale label spaces. We construct our realistic benchmark using two real-world databases. Patient data is sourced from **MIMIC-IV** (Johnson et al., 2023), a large, de-identified electronic health record (EHR) database. We extract adult patient admissions (≥ 18 years) with sufficient prescription history. Each patient context $\mathbf{x}$ aggregates demographics (age, sex), diagnosis codes (ICD-9/ICD-10), and their recorded prescriptions. The ground truth for risk is established using **DrugBank** (Wishart et al., 2018), a comprehensive drug database. We restrict the candidate drugs to those present in both databases, yielding a final label space $\mathcal{Y}$ of 3,421 unique drugs. Patient contexts are split by unique patient ID into calibration (2,000), validation (2,000), and test (2,500) sets, ensuring no patient appears in multiple sets. This creates a long-tail distribution orders of magnitude more complex than standard benchmarks (e.g., CIFAR-100), providing a rigorous testbed for validating risk control stability under extreme sparsity and class imbalance.

Predictor Construction. The base predictor F is a deterministic `FixedRealisticDrugPredictor`, designed to generate clinically plausible yet reproducible probability distributions $P(\cdot|\mathbf{x})$. Its construction follows two steps. First,

we compute the empirical prescription frequencies from the training portion of MIMIC-IV for each diagnostic context to form the backbone of a categorical distribution. Second, to avoid assigning zero probability to clinically essential first-line drugs that may be underrepresented in the data, we apply a Dirichlet-style smoothing anchored by guideline-informed lower bounds (e.g., ensuring lisinopril always receives a non-negligible probability for hypertensive patients). This rule-based design provides a stable testbed where differences in outcomes can be attributed directly to the conformal procedure under study.

F.2. The Risk-Alignment Protocol

Motivation. A direct comparison between CNCRC and baselines like Standard CP is challenging because they control different quantities (R_0 vs. α). To ensure a fair comparison, we must evaluate all methods at the same level of operational risk.

The Three-Step Protocol. We use a three-step protocol involving calibration, validation, and test sets to align all methods to a common risk target R_0.

1. **Calibrate on $\mathcal{D}_{\text{cal}}$.** For CNCRC, we compute its risk-weighted scores and set its conformal threshold q directly, using an internal $\alpha_* = R_0/C_{max}$. For all baselines (Standard CP, Cost-Aware CP, CRC), we calibrate their internal thresholds in the usual way across a pre-defined grid of possible α values.
2. **Select Operating Points on $\mathcal{D}_{\text{val}}$.** For each baseline, we evaluate its performance for every α in the grid on the validation set. We then compute the realized non-coverage risk, $\widehat{R}_{NC}^{(\text{val})}$, for each α. We select the specific $\alpha^\dagger$ that results in a validation risk closest to our target R_0. CNCRC requires no tuning in this step as its threshold q is already fixed.
3. **Report on $\mathcal{D}_{\text{test}}$.** With the parameters now frozen for all methods (q for CNCRC and $\alpha^\dagger$ for each baseline), we perform a final evaluation on the held-out test set and report all metrics.

This protocol ensures that any observed differences in $\widehat{\text{AmbCost}}$ or other efficiency metrics are due to the methods' inherent properties, rather than a mismatch in their operating points.