Conformal Tail Risk Control for Large Language Model Alignment

Catherine Chen Stanford University
Jingyan Shen New York University
Zhun Deng UNC Chapel Hill
Lihua Lei Stanford University

Abstract

Recent developments in large language models (LLMs) have led to their widespread usage for various tasks. The prevalence of LLMs in society implores the assurance on the reliability of their performance. In particular, risk-sensitive applications demand meticulous attention to unexpectedly poor outcomes, i.e., tail events, for instance, toxic answers, humiliating language, and offensive outputs. Due to the costly nature of acquiring human annotations, general-purpose scoring models have been created to automate the process of quantifying these tail events. This phenomenon introduces potential human-machine misalignment between the respective scoring mechanisms. In this work, we present a lightweight calibration framework for blackbox models that ensures the alignment of humans and machines with provable guarantees. Our framework provides a rigorous approach to controlling any distortion risk measure that is characterized by a weighted average of quantiles of the loss incurred by the LLM with high confidence. The theoretical foundation of our method relies on the connection between conformal risk control and a traditional family of statistics, i.e., L-statistics. To demonstrate the utility of our framework, we conduct comprehensive experiments that address the issue of human-machine misalignment.

1 Introduction

Large Language Models (LLMs) have proven to be pervasive in society with applications across various settings, including those of high sensitivity. While LLMs generally perform quite well, there remains a low probability associated with the event that the models generate undesirable and even catastrophic outputs.

Quantitative notions of disutility, such as toxicity, are typically based on human annotations that are costly to acquire. To reduce labor cost, general purpose models, for example, Detoxify, Hanu and Unitary team [2020], have been created to automatically generate disutility measures for LLM outputs. Given the existence of both human and machine assessments, a common issue is the misalignment of the two metrics, which might be caused by distribution shift of human opinions when deploying the model. This is further complicated by the lack of an inherent scale that governs machine scores. While many techniques, such as, Reinforcement Learning from Human Feedback (RLHF), have been proposed to improve alignment, these approaches typically do not provide guarantees regarding the alignment of machine-generated scores. Even when such guarantees exist, they often rely on assumptions that are hard to defend and require computationally intensive model refitting [Christiano et al., 2017, Ziegler et al., 2019, Casper et al., 2023].

In this work, we address the issue of misalignment through the lens of risk control. We treat the human-annotated disutility score as the ground-truth risk measure and calibrate the raw outputs generated by the LLM to control certain functionals of the risk distribution at a pre-specified level.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

Our method does not involve model refitting and provides finite-sample guarantees of risk control under no assumptions about the LLM or the underlying data generation process.

The proposed framework substantially generalizes the existing literature in conformal risk control Bates et al. [2021], Angelopoulos et al. [2021, 2023], which only holds for traditional risk measures characterized by the expectation of a loss function. These risk measures are not suitable for tail risks associated with low probability events. By contrast, other works only consider risk measures that are quantiles of a loss function on the human scores of the outputs [Mohri and Hashimoto, 2024, Cherian et al., 2024, Quach et al., 2024]. A better suited metric to account for information from the more extreme quantiles is the Conditional Value-at-Risk (CVaR) Rockafellar and Uryasev [2000], which measures the average of a range of upper quantiles and has been considered by Snell et al. [2022] and Zollo et al. [2023]. In this work, we explore how distortion risk control can be applied to align LLMs with respect to any disutility metric and leverage L-estimators van der Vaart [1998] to achieve finite-sample control of any distortion risk measure.

2 Problem setup

An LLM produces a response $y(x) \in \mathcal{Y}$ for any given user prompt $x \in \mathcal{X}$ from a distribution $p(y \mid x)$. The response $y(x)$ could be an answer to a question or a response to a comment made by the user. To evaluate the disutility of $y(x)$, human-annotators are enlisted to rate different aspects of $y(x)$, for instance, misinformation and toxicity. Let $r(y(x))$ denote the human rating of $y(x)$, which is generally random due to cognitive uncertainty. Throughout the paper we assume that $r(y(x)) = 0$ when the LLM declines to respond. See an example of a disutility metric in Appendix F.1.

In most applications, one can leverage historical data to train a model that estimates the human-rated disutility. For example, Detoxify [Hanu and Unitary team, 2020] is a machine learning model that assesses the toxicity of responses. Denote $r_m(y(x))$ as the machine-generated disutility score for a response y. Note that $r_m(y(x))$ typically differs from $r(y(x))$ and may even lack monotonicity with respect to $r(y(x))$. Although machine ratings are inexpensive and scalable, the misalignment, or lack of rank preservation between the machine and human ratings diminishes its reliability. Moreover, it is often hard to interpret the scale of the score and choose the right cut-off to decide whether the generated data should be accepted, especially when applied to different contexts.

Given any generative model or sampler $p(y \mid x)$, an aligned model produces an output $\tilde{y}(x)$ in a way such that the overall human-rated disutility, $r(\tilde{y}(x))$, is minimized. We call $\tilde{y}(x)$ a calibrated model.

Let x be a random draw from the population of prompts of interest, and $F_{r(\tilde{y}(x))}$ denote the distribution of $r(\tilde{y}(x))$ over x, which depends on the randomness of $\tilde{y}(\cdot)$, and cognitive uncertainty of $r(\cdot)$. To aggregate over different prompts and integrate out the randomness of responses and cognitive uncertainty, we can define a summary measure $R\left(F_{r(\tilde{y}(x))}\right)$ where $F_{r(\tilde{y}(x))}$ denotes the cumulative distribution function (CDF) of $r(\tilde{y}(x))$ with x being a draw from a population of prompts, and $R(\cdot)$ being a functional that maps any distribution to a non-negative number. We denote $R\left(F_{r(\tilde{y}(x))}\right)$ using $R(F)$ as a notational shorthand. As an example, $R(F)$ can be chosen as the mean disutility.

To achieve alignment, we aim to control the risk $R\left(F_{r(\tilde{y}(x))}\right)$ at a pre-specified level α. This objective is different from existing practices (e.g., RLHF) that target the risk associated with the machine rather than the human disutility scores. This is a challenging task because the human rating function $r(\cdot)$ operates as a black box.

Note that any risk level α can be achieved by abstaining from responding to any prompts. Clearly, this should be avoided. As will be seen later, the deployment cost of our calibrated LLM increases as α decreases and could grow to infinity as α approaches 0, if no abstention is allowed. As a result, though being conservative may appear innocuous, it is unnecessarily costly. To minimize deployment costs, we would want $R\left(F_{r(\tilde{y}(x))}\right)$ to be as close to α as possible.

3 Distortion risk control via L-statistics

3.1 Theoretical setting

Suppose we sample n prompts $x_1, \ldots, x_n$ i.i.d. from a distribution. For each prompt, we generate the candidate set $\mathcal{C}(x_i)$ as described in Section D.1, then recruit human raters to score all responses in the

set. Following the procedure in Section D.1, we can obtain a dataset $\mathcal{D}$ that includes $(x_i, \{r_\lambda(x_i)) : \lambda \in \Lambda\})$ for $i = 1, \ldots, n$. We assume that the machine disutility score model is pretrained and independent of our dataset $\mathcal{D}$. Then the data points $(x_i, \{r_\lambda(x_i)) : \lambda \in \Lambda\})$ remain i.i.d. Throughout the rest of the section we denote a generic draw from the prompt distribution by x.

For each $\lambda \in \Lambda$, we use the shorthand notation $R_\psi(\lambda)$ for $R_\psi(F_{r_\lambda(x)})$, where R_ψ is the distortion risk measure defined in (4) in Appendix A.2 with a user-chosen weight measure ψ. Our goal is to learn $\hat{\lambda}$ from $\mathcal{D}$ such that

$$\mathbb{P}_{\mathcal{D}}(R_\psi(\hat{\lambda}) \leq \alpha) \geq 1 - \delta,$$

for some pre-specified (α, δ). Above, $\mathbb{P}_{\mathcal{D}}$ accounts for the randomness in $\mathcal{D}$. The parameter α represents the target risk level and $1 - \delta$ corresponds to the confidence level.

3.2 Distortion risk measures as L-statistics

Considering that the majority of generated data is normal, traditional risk measures, i.e., the expectation of a loss function, may fail to capture the disutility as these events are only manifested in the tail. Hence, we choose $R(F)$ to be a *distortion risk measure* [e.g. Balbás et al., 2009, Snell et al., 2022], defined as a weighted average of quantiles.

In our setting, fixing $\lambda \in \Lambda$, let $r_{\lambda,(1)} \leq r_{\lambda,(2)} \leq \ldots \leq r_{\lambda,(n)}$ denote the ordered statistics of $(r_\lambda(x_1), \ldots, r_\lambda(x_n))$. Furthermore, let F_λ and $\hat{F}_{n,\lambda}$ be the true and empirical distributions of $(r_\lambda(x_1), \ldots, r_\lambda(x_n))$, respectively. For a given distortion risk measure, the plug-in estimator that replaces the true distribution F by the empirical distribution $\hat{F}_{n,\lambda}$ is

$$\hat{R}_\psi(\lambda) = R_\psi(\hat{F}_{n,\lambda}) = \int \hat{F}^{-1}_{n,\lambda}(p) d\psi(p). \tag{1}$$

L-statistics refers to the class of estimators expressed as linear combinations of order statistics originating from Mosteller [1946]. Notable examples include the sample quantile, the trimmed mean, and the winsorized mean [Tukey, 1962]. By definition, $\hat{F}^{-1}_{n,\lambda}(p) = r_{\lambda,(i)}$ for any $p \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$, thus it can be written as an L-statistic

$$\hat{R}_\psi(\lambda) = \sum_{i=1}^{n} \left\{ \psi\left(\frac{i}{n}\right) - \psi\left(\frac{i-1}{n}\right) \right\} r_{\lambda,(i)}.$$

Theorem 22.3 of van der Vaart [1998] describes the asymptotic normality of L-statistics. Given this result, we show that $\hat{R}_\psi(\lambda)$ is an asymptotically normal estimator of $R_\psi(\lambda)$ for any fixed $\lambda \in \Lambda$ with a consistent variance estimator in Theorem B.1, and the proof in Appendix C. We assume the boundedness of $r_\lambda(x_i)$ for simplicity, though this can be relaxed with [Gardiner and Kumar Sen, 1979, Stigler, 1974].

3.3 Conformal distortion risk control via L-statistics

By design, $R_\psi(\lambda)$ is monotonic in λ. This allows us to apply the method of Bates et al. [2021] to achieve risk control by inverting a pointwise upper confidence bound (UCB). Specifically, we choose

$$\hat{\lambda} = \max\{\lambda \in \Lambda : \hat{R}^+_\psi(\lambda') \leq \alpha, \, \forall \lambda' \leq \lambda\}, \tag{2}$$

where

$$\hat{R}^+_\psi(\lambda) = \hat{R}_\psi(\lambda) + z_{1-\delta} \cdot \hat{\sigma}(\lambda),$$

and $z_{1-\delta}$ is the $(1-\delta)$-th quantile of the standard normal distribution. In practice, we discretize Λ to avoid checking the condition $\hat{R}^+_\psi(\lambda) \leq \alpha$ for infinitely many points. Our procedure is outlined in Algorithm 2 of Appendix E, and illustrated in Fig. 4 of Appendix .

By Theorem B.1, $\hat{R}^+_\psi(\lambda)$ is an asymptotic $(1-\delta)$ UCB. Although Bates et al. [2021] focus on traditional risk measures that are expressed as expected loss, we can easily extend their result (Theorem 6) to distortion risk measures via Theorem B.3. Remarkably, this result does not involve any assumption on the underlying LLM $p(y \mid x)$ or machine disutility score model $r_m(x)$. In particular, it allows the machine ratings to be arbitrarily misaligned with human ratings.

3.4 Alternative (conservative) approaches for distortion risk control

Another strategy to construct pointwise UCBs for $R_\psi(\lambda)$ is to replace all quantiles by their confidence envelopes. Specifically, if we have statistics $\{\hat{q}_p^+(\lambda) : p \in [0,1]\}$ such that

$$\mathbb{P}(q_p(\lambda) \leq \hat{q}_p^+(\lambda),\ \forall p \in [0,1]) \geq 1-\delta, \tag{3}$$

then $\int_0^1 \hat{q}_p^+(\lambda) d\psi(p)$ is a valid $(1-\delta)$ UCB for $R_\psi(\lambda)$. Following Snell et al. [2022], we consider two confidence envelopes based on the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality and Berk-Jones (BJ) statistics, respectively. The details of these statistics are outlined in Appendix E.1.

3.5 Results

We illustrate the performance of our algorithm via experimentation with details outlined in Appendix F. Here, we highlight the deployment results of the Distortion Risk Control procedure deployed on the LLAMA-2-7B-HF model.

Human annotations are the primary contributors to the calibration cost in our framework - our definition of deployment cost is described in Appendix D.5. Nevertheless, we demonstrate that our method retains statistical efficiency even with smaller calibration sets, such as, $|\mathcal{D}| \in \{50, 100, 200, 1000\}$.

Below, we highlight the results for CVaR_β, where β = 0.5, with risk controlled at level $\alpha = 0.25$ with confidence $1-\delta = 0.95$, and Spearman correlation between the human and machine toxicity scores at $\rho = 0.57$. We construct confidence bands by taking the mean estimate $\pm$ one standard error estimated from the results across 15 independent experiments. We use CDRC as a notational shorthand for Conformal Distortion Risk Control. Thus CDRC-L, CDRC-DKW, and CDRC-BJ represent CDRC via L-statistics, the DKW inequality, and the Berk Jones statistics, respectively.

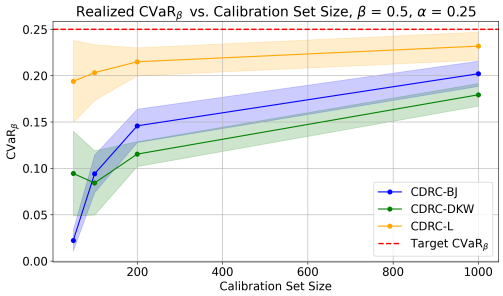

Figure 1: **LLaMA2-7B: Realized CVaR_β vs. calibration set size**.

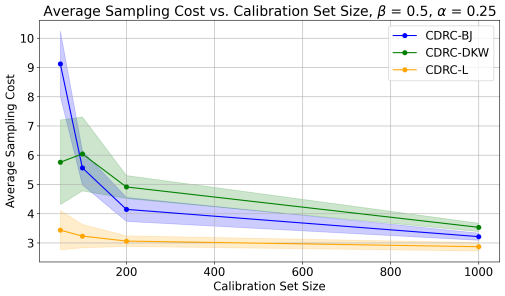

Figure 2: **LLaMA2-7B: Average sampling cost for CVaR_β control vs. calibration set size**.

From Fig. 1, it is evident that CDRC-L is consistently the least conservative among all methods, while still maintaining the guaranteed CVaR_β control. Furthermore, from Fig. 2, we observe that CDRC-L is consistently the most cost effective among all methods. In particular, it is significantly less expensive in comparison to other methods in settings with small calibration set sizes. We show qualitatively equivalent results for VaR_β in Appendix H.

We present the realized risk and average cost analysis for CVaR_β and VaR_β control with $\rho \in \{0.57, 0.68, 0.78\}$ in Appendix G and Appendix H, respectively. The comparison between CDRC-L, CDRC-DKW, and CDRC-BJ are qualitatively similar. Notably, the realized CVaR_β and VaR_β of CDRC-L is extremely close to the target level with low sampling cost, demonstrating that it is not conservative while remaining the most cost effect among all methods. Furthermore, experiments outlining the relationship between cost and misalignment, and comparison to best-of-N are described in Appendix G.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *CoRR*, abs/2110.01052,

2021. URL https://arxiv.org/abs/2110.01052.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2023. URL https://arxiv.org/abs/2208.02814.

Alejandro Balbás, José Garrido, and Silvia Mayoral. Properties of distortion risk measures. *Methodology and Computing in Applied Probability*, 11(3):385–399, 2009.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets, 2021.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Stephen Casper, S Zhang, Ari Holtzman, Julia Kreutzer, Chenguang Zhu, Marc'Aurelio Ranzato, and Nisan Stiennon. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods, 2024. URL https://arxiv.org/abs/2406.09714.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017.

cjadams, Daniel Borkan andinversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. Jigsaw unintended bias in toxicity classification, 2019. URL https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification.

Miklós Csörgő and Pál Révész. Strong approximations of the quantile process. *The Annals of Statistics*, pages 882–894, 1978.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Joseph C Gardiner and Pranab Kumar Sen. Asymptotic normality of a variance estimator of a linear combination of a function of order statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 50(2):205–221, 1979.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Wassilij Hoeffding. Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und Instituts fur Angewandte Mathematik der Universitat Berlin*, 5:181–233, 1940.

J.C. Kiefer. Deviations between the sample quantile process and the sample df. *Non-parametric Techniques in Statistical Inference*, pages 299–319, 1970.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees, 2024. URL https://arxiv.org/abs/2402.10978.

Amit Moscovich and Boaz Nadler. Fast calculation of boundary crossing probabilities for poisson processes. *Statistics & Probability Letters*, 123:177–182, 2017.

F Mosteller. On some useful" inefficient" statistics. *Annals of Mathematical Statisitcs*, 17:377–408, 1946.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling, 2024. URL https://arxiv.org/abs/2306.10193.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk, 2000.

Jake C. Snell, Thomas P. Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions, 2022. URL https://arxiv.org/abs/2212.13629.

Stephen M Stigler. Linear functions of order statistics with smooth weight functions. *The Annals of Statistics*, pages 676–693, 1974.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

John W Tukey. The future of data analysis. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 408–452. Springer, 1962.

Aad van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models, 2023.

Supplementary materials

In the supplementary materials, we provide implementation details, additional experimental results, and formalized proofs. Code repository can be found at https://github.com/jy-evangeline/DRC.

A Background

A.1 Related works and contribution

Quantile Risk Control. Traditionally, conformal risk control is characterized by bounding the expected loss of a given predictor to obtain provable guarantees [Bates et al., 2021, Angelopoulos et al., 2021]. Recent work has extended the class of loss functions to a wider class of functions defined by the loss for the quantiles of the data distribution, i.e., distortion risks Snell et al. [2022]. Existing methods rely on replacing all quantiles by their confidence envelope to form an *upper confidence bound* (UCB) on the true risk, as described in Bates et al. [2021]. These procedures are presented in more detail in 3.4. In our work, we directly estimate any given distortion risk measure via *L-statistics* to derive tight risk control bounds.

Connection to Best-of-N. Our work differs from inference-time alignment strategies, such as, best-of-N, a fixed-sample inference-time heuristic. In contrast, our method is an adaptive, risk-controlling strategy. In particular, best-of-N uses a fixed generation count N, meaning that it always samples N responses for each prompt. Conversely, our framework uses an adaptive number of samples, N_x, depending on the prompt and desired risk level. For example, when a prompt is non-toxic, best-of-N generates N responses, while our method may only require one, yielding a lower average inference cost. We illustrate this difference between the two methods in G. Furthermore, best-of-N implicitly assumes that the disutility function in the LLM framework is well-aligned with the human user. Not only does our method not make such assumptions, but we also extend the problem of alignment by introducing and tuning a parameter that effectively controls the human distuility of an LLM generated response.

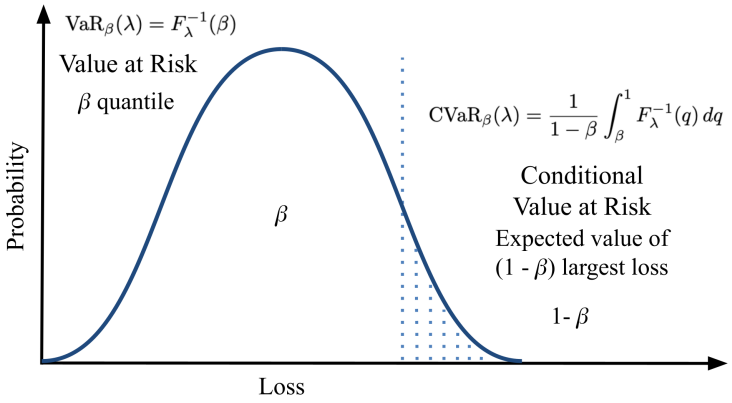

Figure 3: **Examples of distortion risk measures: Value-at-Risk (VaR_β) and Conditional Value-at-Risk (CVaR_β).**

A.2 Distortion risk measures and L-statistics

We choose $R(F)$ to be a *distortion risk measure* [e.g. Balbás et al., 2009, Snell et al., 2022], defined as a weighted average of quantiles,

$$R_\psi(F) := \int_0^1 F^{-1}(p)\, d\psi \tag{4}$$

where $F^{-1}(p) \triangleq \inf\{x : F(x) \geq p\}$ denotes the p-th quantile of F, and $\psi(\cdot)$ is a weighting measure such that $\psi(p) \geq 0$ and $\int_0^1 d\psi(p) = 1$. An example of a distortion risk measure is the widely-used CVaR_β Rockafellar and Uryasev [2000] where ψ is the uniform measure on $[\beta, 1]$. Other examples include mean (with ψ the uniform measure on $[0, 1]$) and β-th quantile, also known as Value-at-Risk, for any β (with ψ the point mass at p). Examples of distortion risk measures are shown in (Fig. 3). We also want to remark that expected mean and quantiles are also special cases of distortion risk measures.

Notation

A summary of the notation used in the paper is shown in Table 1.

B Theoretical results

This section highlights the theortical results of the paper.

Theorem B.1. *Assume $r_\lambda(x) \in [a, b]$ almost surely for some $-\infty < a < b < \infty$, F_λ is continuous and strictly increasing. Further, assume that $\psi(y) = \int_0^y \psi'(z)dz$ for some ψ' that is bounded and continuous at $F_\lambda(r)$ for Lebesgue almost-every r. Then,*

$$\frac{\sqrt{n}(\hat{R}_\psi(\lambda) - R_\psi(\lambda))}{\hat{\sigma}(\lambda)} \xrightarrow{d} N(0, 1),$$

where

$$\hat{\sigma}^2(\lambda) = \int\int \psi'(\hat{F}_{n,\lambda}(r))\psi'(\hat{F}_{n,\lambda}(\tilde{r}))D_\lambda(r, \tilde{r})\, dr\, d\tilde{r},$$

with $\hat{F}_{n,\lambda}$ being the empirical distribution of $r_\lambda(x_1), \ldots, r_\lambda(x_n)$ and

$$D_\lambda(r, \tilde{r}) = \hat{F}_{n,\lambda}(r \wedge \tilde{r}) - \hat{F}_{n,\lambda}(r)\hat{F}_{n,\lambda}(\tilde{r}).$$

Equivalently,

$$\hat{\sigma}^2(\lambda) = \frac{1}{n^2}\sum_{i,j=1}^{n} \psi'\left(\frac{i}{n}\right)\psi'\left(\frac{j}{n}\right)\left(\frac{i \wedge j}{n} - \frac{ij}{n^2}\right). \tag{5}$$

Table 1: Summary of notation used in the paper.

Symbol	Meaning
$\mathcal{X}$	User prompt space.
$\mathcal{Y}$	LLM output space.
x	User prompt drawn from $\mathcal{X}$.
$y(x)$	LLM response given a user prompt $x \in \mathcal{X}$.
$\tilde{y}(x)$	Calibrated LLM response given a user prompt $x \in \mathcal{X}$.
$p(x \mid y)$	Data generating process of LLM.
$\mathcal{C}(x)$	Set of uncalibrated responses generated by an LLM, where $\lvert\mathcal{C}(x)\rvert = N$ for a given prompt x.
$\mathcal{C}_\lambda(x)$	Set of responses y generated by an LLM satisfying $r_m(y) < \lambda$ for a given prompt x.
$r(y)$	Human-generated disutility score of response $y(x)$.
$r_m(y)$	Machine-generated disutility score of response $y(x)$.
$r_\lambda(x)$	Disutility score for $\mathcal{C}_\lambda(x)$ as the worst human rating, i.e. $r_\lambda(x) = \max_{y \in \mathcal{C}_\lambda(x)} r(y)$.
$F_{r(\tilde{y}(x))}$	Cumulative distribution function (CDF) of $r(\tilde{y}(x))$ over x.
$R(F_{r(\tilde{y}(x))})$	Summary measure of $F_{r(\tilde{y}(x))}$, where $R(\cdot)$ is a functional that maps any distribution to a non-negative number. Denote as $R(F)$ as a shorthand.
Λ	Range of machine scores with $\Lambda = [\lambda_{\min}, \lambda_{\max}]$
$\mathcal{D}$	Dataset of $(x_i, \{r_\lambda(x_i) : \lambda \in \Lambda\})$ for n i.i.d. prompts $x_1, \ldots, x_n$.
$R_\psi(F_{r(\tilde{y}(x))})$	Distortion risk measure, where $\psi(\cdot)$ is a weighting measure such that $\psi(p) \geq 0$, and $\int_0^1 d\psi(p) = 1$. Denote as $R_\psi(\lambda)$ as a shorthand.
F_λ	True distribution of $(r_\lambda(x_1), \ldots, r_\lambda(x_n))$.
$\hat{F}_{n,\lambda}$	Empirical distribution of $(r_\lambda(x_1), \ldots, r_\lambda(x_n))$.
$\hat{R}_\psi(\lambda)$	Empirical distortion risk measure of $(r_\lambda(x_1), \ldots, r_\lambda(x_n))$ with $\hat{R}_\psi(\lambda) = \hat{R}_\psi(\hat{F}_{n,\lambda}) = \int \hat{F}_{n,\lambda}^{-1}(p) d\psi(p)$.
α	Prespecified risk control level.
$1 - \delta$	Prespecified confidence level.

Among all distortion risk measures, CVaR_β is the most interpretable and widely-used metric. For CVaR_β, we can find a much simpler variance estimator.

Corollary B.2. *For CVaR_β with $\psi(p) = \max\{p - \beta, 0\}/(1 - \beta)$, Theorem B.1 holds with*

$$\hat{\sigma}^2(\lambda) = \frac{1}{(1-\beta)^2} \widehat{\mathrm{Var}} \left(\{\max\{r_\lambda(x_i), r_{\lambda,(\lceil n\beta \rceil)}\}_{i=1}^n \} \right),$$

where $\widehat{\mathrm{Var}}$ denotes the sample variance.

Another important example is VaR_β. While the VaR is not a distortion risk measure with a differentiable ψ, we develop parallel theory in Appendix H based on the asymptotic theory of empirical quantiles. Unlike CVaR_β, the asymptotic variance depends on the density F'_λ of r_i and hence harder to estimate. We apply the bootstrap technique [Efron and Tibshirani, 1994] instead to estimate $\hat{\sigma}^2(\lambda)$.

Theorem B.3. *Let $\hat{\lambda}$ be defined in* (2)*. Assume $R_\psi(\lambda)$ is continuous and strictly increasing. In the same setting as Theorem B.1*

$$\liminf_{n \to \infty} \mathbb{P}_{\mathcal{D}} \left(R_\psi \left(\hat{\lambda} \right) \leq \alpha \right) \geq 1 - \delta. \tag{6}$$

As a consequence, for any selection mechanism that picks $\tilde{y}(x)$ from $\mathcal{C}_{\hat{\lambda}}(x)$,

$$\liminf_{n \to \infty} \mathbb{P}_{\mathcal{D}} \left(R_\psi \left(F_{r(\tilde{y}(x))} \right) \leq \alpha \right) \geq 1 - \delta.$$

C Proofs

C.1 Proof of theorem C.1

We decompose the proof into two parts.

Theorem C.1 (Asymptotic normality of L-statistics)**.** *Let f_λ, F_λ denote the PDF and CDF of the score r_i, respectively, and $\hat{F}_{n,\lambda}$ denote the empirical CDF of the scores $r_1, \ldots, r_n$. Further, let*

$$R_\psi(\lambda) = \int_0^1 F_\lambda^{-1}(p) d\psi(p) \quad \textit{and} \quad \hat{R}_\psi(\lambda) = \int \hat{F}_{n,\lambda}^{-1}(p) d\psi(p),$$

for any measure ψ that may not have a density (e.g., ψ can be a point mass at $p_0 \in (0,1)$). Suppose

1. *$r_i \in [a,b]$ for some $-\infty < a < b < \infty$ almost surely, with $\inf_{r\in[a,b]} f_\lambda(r) > 0$, and $\sup_{r\in[a,b]} f'_\lambda(r) < \infty$.*
2. *$\int \frac{d\psi(t)}{f_\lambda(F_\lambda^{-1}(t))} < \infty$.*

then $\sqrt{n}(\hat{R}_\psi(\lambda) - R_\psi(\lambda)) \xrightarrow{d} N(0, \sigma^2(\lambda))$ where

$$\sigma^2(\lambda) = \int_0^1 \int_0^1 \frac{p \wedge p' - p \cdot p'}{f_\lambda(F_\lambda^{-1}(p)) f_\lambda(F_\lambda^{-1}(p'))} d\psi(p) d\psi(p') < \infty.$$

Proof. By the Bahadur-Kiefer representation Kiefer [1970] (see also Theorem E of Csörgő and Révész [1978]) of sample quantiles we can write

$$\begin{aligned}
\Delta_n &\triangleq \sup_{p\in[0,1]} |(F_{n,\lambda}^{-1}(p) - F_\lambda^{-1}(p)) f_\lambda(F_\lambda^{-1}(p)) - (F_{n,\lambda}(F_\lambda^{-1}(p)) - p)| \\
&= O_p\left(\frac{\log(n)^{1/2}(\log\log n)^{1/4}}{n^{3/4}}\right) = O_p\left(\frac{\log(n)}{n^{3/4}}\right),
\end{aligned}$$

equivalently,

$$\left|(\hat{F}_{n,\lambda}^{-1}(p) - F_\lambda^{-1}(p)) - \frac{F_{n,\lambda}(F_\lambda^{-1}(p)) - p}{f_\lambda(F_\lambda^{-1}(p))}\right| = \frac{\Delta_n}{f_\lambda(F_\lambda^{-1}(p))}.$$

Given this, we have that

$$\begin{aligned}
\hat{R}_\psi(\lambda) - R_\psi(\lambda) &= \int_0^1 (\hat{F}_{n,\lambda}^{-1}(p) - F_\lambda^{-1}(p)) d\psi(p) \\
&= \int_0^1 \frac{F_{n,\lambda}(F_\lambda^{-1}(p)) - p}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) + \delta_n
\end{aligned}$$

where

$$|\delta_n| \le \Delta_n \cdot \int_0^1 \frac{1}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) = O_p\left(\frac{\log(n)}{n^{3/4}}\right)$$

holds by *assumption 2*.

Thus we have that

$$\begin{aligned}
\sqrt{n}(\hat{R}_\psi(\lambda) - R_\psi(\lambda)) &= \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^n \int_0^1 \frac{\mathbb{1}\{r_i \le F_\lambda^{-1}(p)\} - p}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) + o_p(1) \\
&\triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n S(r_i) + o_p(1)
\end{aligned}$$

By *assumption 2*,

$$S(r_i) = \int_0^1 \frac{\mathbb{1}\{r_i \le F_\lambda^{-1}(p)\} - p}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) \le \int_0^1 \frac{1}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) < \infty.$$

It remains to prove $\mathbb{E}[S(r_i)] = 0$ and $\mathrm{Var}(S(r_i)) = \mathbb{E}[S^2(r_i)] < \infty$.

To study the mean, note that by *assumption 1*, $F_\lambda^{-1}(p)$ is uniquely defined and $\mathbb{P}(r_i \le F_\lambda^{-1}(p)) = p$. Thus,

$$\mathbb{E}[S(r_i)] = \int_0^1 \frac{\mathbb{E}[\mathbb{1}\{r_i \le F_\lambda^{-1}(p)\} - p]}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) = 0$$

To study the variance, observe that the second moment can be expressed as follows

$$\begin{aligned}
\mathbb{E}[S^2(r_i)] &= \mathbb{E}\left[\left(\int_0^1 \frac{\mathbb{1}\{r_i \le F_\lambda^{-1}(p)\} - p}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p)\right)^2\right] \\
&= \mathbb{E}\left[\int_0^1 \frac{\mathbb{1}\{r_i \le F_\lambda^{-1}(p)\} - p}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) \cdot \int_0^1 \frac{\mathbb{1}\{r_i \le F_\lambda^{-1}(p')\} - p'}{f_\lambda(F_\lambda^{-1}(p'))} d\psi(p')\right] \\
&= \int_0^1 \int_0^1 \frac{\mathbb{E}\left[(\mathbb{1}\{r_i \le F_\lambda^{-1}(p)\} - p) \cdot (\mathbb{1}\{r_i \le F_\lambda^{-1}(p')\} - p')\right]}{f_\lambda(F_\lambda^{-1}(p)) f_\lambda(F_\lambda^{-1}(p'))} d\psi(p) d\psi(p')
\end{aligned}$$

Let $B(p) = \mathbb{1}\{r_i \le F_\lambda^{-1}(p)\}$. Notice that $\mathbb{E}[B(p)] = p$, thus

$$\begin{aligned}
\mathbb{E}\left[(B(p) - p) \cdot (B(p') - p')\right] &= \mathrm{Cov}(B(p), B(p')) \\
&= \mathbb{E}\left[B(p) \cdot B(p')\right] - p \cdot p' \\
&= \begin{cases} p - p \cdot p' & , \quad \text{if } p \le p' \\ p' - p \cdot p' & , \quad \text{if } p > p' \end{cases} \\
&= p \wedge p' - p \cdot p'
\end{aligned}$$

Therefore,

$$\mathrm{Var}(S(r_i)) = \int_0^1 \int_0^1 \frac{p \wedge p' - p \cdot p'}{f_\lambda(F_\lambda^{-1}(p)) f_\lambda(F_\lambda^{-1}(p'))} d\psi(p) d\psi(p')$$

By *assumption 2*,

$$\mathrm{Var}(S(r_i)) \le \int_0^1 \int_0^1 \frac{1}{f_\lambda(F_\lambda^{-1}(p)) f_\lambda(F_\lambda^{-1}(p'))} d\psi(p) d\psi(p') = \left|\int_0^1 \frac{1}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p)\right|^2 < \infty.$$

□

Theorem C.2 (Consistent variance estimate for L-Statistics)**.** *In the same setting as Theorem C.1, assume the assumption 1 holds and that $\psi(y) = \int_0^y \psi'(z)dz$ for some ψ' that is bounded and continuous at $F_\lambda(r)$ for Lebesgue almost-every r.*

Let

$$\hat{\sigma}_n^2(\lambda) = \int_a^b \int_a^b \psi'(\hat{F}_{n,\lambda}(r)) \psi'(\hat{F}_{n,\lambda}(\tilde{r})) (\hat{F}_{n,\lambda}(r \wedge \tilde{r}) - \hat{F}_{n,\lambda}(r) \hat{F}_{n,\lambda}(\tilde{r}))\, dr\, d\tilde{r}.$$

If ψ' is bounded, then $\hat{\sigma}_n^2(\lambda) \xrightarrow{a.s.} \sigma^2(\lambda)$ as $n \to \infty$.

Proof. We first verify *assumption 2*. Take $p = F_\lambda(x)$,

$$\begin{aligned}
\int_0^1 \frac{1}{f_\lambda(F_\lambda^{-1}(p))} d\psi(p) &= \int \frac{\psi'(p)}{f_\lambda(F_\lambda^{-1}(p))} dp \\
&= \int_a^b \frac{\psi'(F_\lambda(x))}{f_\lambda(x)} dF_\lambda(x) \\
&= \int_a^b \psi'(F_\lambda(x)) dx.
\end{aligned}$$

The integral is bounded since ψ' is bounded and a, b are both finite. Let $p = F_\lambda(r)$, and $p' = F_\lambda(\tilde{r})$ in the double integral of $\sigma^2(\lambda)$, then

$$\begin{aligned}\sigma^2(\lambda) &= \int_0^1 \int_0^1 \frac{p \wedge p' - p \cdot p'}{f_\lambda(F_\lambda^{-1}(p)) f_\lambda(F_\lambda^{-1}(p'))} d\psi(p) d\psi(p') \\ &= \int_a^b \int_a^b \frac{F_\lambda(r \wedge \tilde{r}) - F_\lambda(r)F_\lambda(\tilde{r})}{\cancel{f_\lambda(r) f_\lambda(\tilde{r})}} \psi'(F_\lambda(r))\psi'(F_\lambda(\tilde{r})) \cdot \cancel{f_\lambda(r) f_\lambda(\tilde{r})}\, dr\, d\tilde{r}.\end{aligned}$$

By the Glivenko-Centelli Theorem, $\hat{F}_{n,\lambda}(r) \xrightarrow{a.s.} F_\lambda(r)$ for every $r \in [a, b]$, thus, for any $r, r' \in [a, b]$,

$$\hat{F}_{n,\lambda}(r \wedge \tilde{r}) - \hat{F}_{n,\lambda}(r)\hat{F}_{n,\lambda}(\tilde{r}) \xrightarrow{a.s.} F_\lambda(r \wedge \tilde{r}) - F_\lambda(r)F_\lambda(\tilde{r}).$$

Furthermore, since ψ' is continuous at $F(r)$ for almost every r under the Lebesgue measure, by the continuous mapping theorem,

$$\psi'(\hat{F}_{n,\lambda}(r)) \xrightarrow{a.s.} \psi'(F_\lambda(r))$$

for almost every r under Lebesque measure. Suppose that ψ' is bounded above by B, then

$$\psi'(\hat{F}_{n,\lambda}(r))\psi'(\hat{F}_{n,\lambda}(\tilde{r}))(\hat{F}_{n,\lambda}(r \wedge \tilde{r}) - \hat{F}_{n,\lambda}(r)\hat{F}_{n,\lambda}(\tilde{r})) \le B^2,$$

thus

$$\hat{\sigma}_n^2(\lambda) \xrightarrow{a.s.} \sigma^2(\lambda)$$

by the Dominated Convergence Theorem.

□

C.2 Proof of corollary B.2

For CVaR_β, $\psi(p) = \max\{p - \beta, 0\}/(1 - \beta)$ and hence $\psi' = I(p \ge \beta)/(1 - \beta)$. By (C.2),

$$\begin{aligned}\sigma^2(\lambda) &= \frac{1}{(1-\beta)^2} \int_a^b \int_a^b (F_\lambda(r \wedge \tilde{r}) - F_\lambda(r)F_\lambda(\tilde{r}))I(F_\lambda(r) \ge \beta)I(F_\lambda(\tilde{r}) \ge \beta) dr d\tilde{r} \\ &= \frac{1}{(1-\beta)^2} \int_{F_\lambda^{-1}(\beta)}^b \int_{F_\lambda^{-1}(\beta)}^b (F_\lambda(r \wedge \tilde{r}) - F_\lambda(r)F_\lambda(\tilde{r})) dr d\tilde{r}.\end{aligned}$$

Let G_λ be the CDF of $r_i' \triangleq \max\{r_i, \beta\}$. Then

$$G_\lambda(r) = F_\lambda(r) I(r \ge F_\lambda^{-1}(\beta)).$$

By Hoeffding's covariance identity [Hoeffding, 1940],

$$\mathrm{Var}(r_i') = \int_a^b \int_a^b (G_\lambda(r \wedge \tilde{r}) - G_\lambda(r)G_\lambda(\tilde{r})) dr d\tilde{r} = \int_{F_\lambda^{-1}(\beta)}^b \int_{F_\lambda^{-1}(\beta)}^b (F_\lambda(r \wedge \tilde{r}) - F_\lambda(r)F_\lambda(\tilde{r})) dr d\tilde{r}.$$

Thus,

$$\sigma^2(\lambda) = \frac{1}{(1-\beta)^2} \mathrm{Var}(r_i').$$

Since r_i' is bounded, the Law of Large number implies that $\mathrm{Var}(r_i')$ can be estimated consistently by the empirical variance of $(r_1', \ldots, r_n')$.

C.3 Proof of theorem B.3

The proof is very similar to Bates et al. [2021], except that they only consider expected risk measures. If $R_\psi(\lambda) \le \alpha$ for all $\lambda \in \Lambda$, then the result obviously holds. Assume $\sup_{\lambda \in \Lambda} R_\psi(\lambda) > \alpha$. Since $R_\psi(\lambda)$ is continuous and strictly increasing, it crosses α exactly once. Let λ^* denote the crossing point, i.e., $R_\psi(\lambda^*) = \alpha$. Then

$$R_\psi(\hat{\lambda}) > \alpha \iff \hat{\lambda} > \lambda^* \implies \hat{R}_\psi^+(\lambda^*) \le \alpha,$$

where the last line is due to the definition of $\hat{\lambda}$. By Theorem B.1,

$$\lim_{n \to \infty} \mathbb{P}(\hat{R}_\psi^+(\lambda^*) < R_\psi(\lambda^*)) = \delta.$$

Since $R_\psi(\lambda^*) = \alpha$, we have

$$\limsup_{n \to \infty} \mathbb{P}(R_\psi(\hat{\lambda}) > \alpha) \ge \limsup_{n \to \infty} \mathbb{P}(\hat{R}_\psi^+(\lambda^*) < R_\psi(\lambda^*)) = \delta.$$

The proof is then completed.

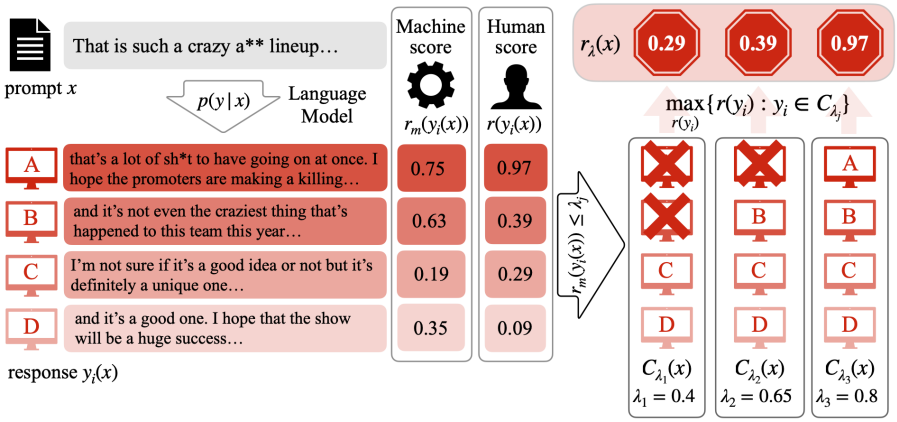

Figure 4: **Illustration of the process to generate $\mathcal{C}(x), \mathcal{C}_\lambda(x)$, and $r_\lambda(x)$.** We sample N responses y_i, $i = 1, \ldots, N$ from the LLM $p(y \mid x)$. Each response is associated with a machine disutility score $r_m(y_i(x))$, and a human-rated disutility score $r(y_i(x))$. To construct $\mathcal{C}_{\lambda_j}(x)$, we keep the responses such that the machine toxicity score satisfies $r_m(y_i(x)) \le \lambda_j$ for each $\lambda_j \in \Lambda$. Finally, we compute the induced score $r_\lambda(x)$ by taking the maximum human disutility score of each $\mathcal{C}_{\lambda_j}(x)$.

D Implementation details

D.1 Augmentation of LLM outputs

If the original output $y(x)$ does not control the risk at α, we refine it by adopting a light-weight calibration approach that amounts to tuning a one-dimensional parameter. Specifically, for every prompt x, we request a response from the LLM multiple times to generate a candidate set $\mathcal{C}(x) = \{y_1(x). \ldots, y_N(x)\}$, where N denotes the set size. To maximize the information content, we follow Quach et al. [2024] in eliminating responses to ensure diversity, quality, and set-confidence; see Appendix D.2 for details.

We generate a nested sequence of subsets of $\mathcal{C}(x)$ that have increasing disutility. Since we are not allowed to collect human ratings on the fly, we use the machine disutility score as a proxy. In particular, for each λ in the range of machine scores Λ, we generate a set

$$\mathcal{C}_\lambda(x) = \{y(x) \in \mathcal{C}(x) : r_m(y) < \lambda\}$$

by using the machine disutility score $r_m(\cdot)$. Without loss of generality, we assume $\Lambda = [\lambda_{\min}, \lambda_{\max}]$. We assign a disutility score $r_\lambda(x)$ for $C_\lambda(x)$ as the worst human rating, i.e.,

$$r_\lambda(x) = \max_{y \in \mathcal{C}_\lambda(x)} r(y).$$

By design, $r_\lambda(x)$ is non-decreasing in λ, a key property that our method leverages. Moreover, $\mathcal{C}_{\lambda_{\min}}(x) = \emptyset$ for any x and hence $r_{\lambda_{\min}}(x) = 0$. The process of generating $\mathcal{C}_\lambda(x)$ and $r_\lambda(x)$ is illustrated in Figure 4.

Notably, $\mathcal{C}_\lambda(x)$ can be computed for any x and λ, because it relies solely on machine disutility scores, whereas $r_\lambda(x)$ is a black-box function that is only available for prompts with collected human ratings. Our goal is to choose $\hat{\lambda}$ based on human-annotated data such that $R(F_{r_{\hat{\lambda}}(x)}) \le \alpha$. Since $r_{\hat{\lambda}}(x)$ is defined as the worst human score in $\mathcal{C}_{\hat{\lambda}}(x)$, we can pick any candidate $\tilde{y}(x) \in \mathcal{C}_{\hat{\lambda}}(x)$ and the resulting risk $R(F_{r(\tilde{y}(x))})$ will be controlled at level α. When $\mathcal{C}_{\hat{\lambda}}(x)$ is empty, the calibrated LLM simply declines to respond. Importantly, the selection can be arbitrary – for example, we could choose the response from $\mathcal{C}_{\hat{\lambda}}(x)$ that has the minimal machine disutility score or maximal information content measured by another metric. To summarize, we reduce the task from retraining a calibrated LLM $\tilde{y}(x)$ with high-dimensional model parameters to searching for a univariate parameter $\hat{\lambda}$.

D.2 Algorithm to generate the candidate set $\mathcal{C}$

Following the original implementation of LLAMA-2-7B-HF model, we set the generation temperature at 0.8 and the top-p parameter at 0.95. We generate candidate response sets for our DRC framework

using Algorithm 1, following Quach et al. [2024]. To ensure quality and diversity of generated candidates, we filter responses by retaining only those with a Perpexity less than 2.61 while ensuring that the ROUGE-L scores between samples in the candidate set is not greater than 0.26. Table 2 presents the percentiles of these metrics across all generated responses. Fig. 5 outlines the cardinality of the sets generated by various combinations of γ.

Algorithm 1 Generation of the candidate set $\mathcal{C}(x)$

Input: input prompt x, set-based confidence function $\mathcal{F}$, text similarity function $\mathcal{S}$, sample quality estimator $\mathcal{Q}$, fixed threshold configuration $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$, sampling budget $k_{\max}$, with conditional output $p_\theta(y \mid x)$ from a generative model.
function CANDIDATESET($x, \mathcal{F}, \mathcal{S}, \mathcal{Q}, \boldsymbol{\gamma}, k_{\max}$)
$\mathcal{C} = \{\}$
for $k = 1$ **to** k_{max} **do**
  $y_k \leftarrow y \sim p_\theta(y \mid x)$ {Sample from generative model}
  if $\mathcal{Q}(x, y_k) < \gamma_1$ and $\max\{\mathcal{S}(y_k, y_j) : y_j \in \mathcal{C}\} > \gamma_2$ **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{y_k\}$ {Quality and similarity control}
  end if
  if $\mathcal{F}(\mathcal{C}) \geq \gamma_3$ **then**
    break{Set-based confidence guarantee}
  end if
end for
return $\mathcal{C}$

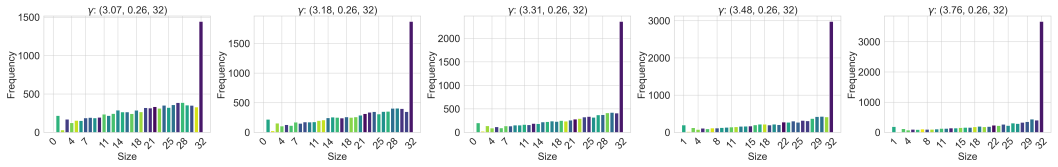

Figure 5: **Size of generated sets under different combinations of hyperparameters**. $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$ refers to (preplexity, similarity, stopping threshold), respectively.

Table 2: Percentiles of different hyperparameters.

Percentile	50	75	80	85	90	95
ROUGE-L	0.26	0.39	0.43	0.47	0.52	0.59
Perplexity	2.61	3.07	3.18	3.31	3.48	3.76

D.3 Detoxify model for $r(\cdot)$ and $r_m(\cdot)$

We use the original Detoxify model as a proxy for human annotator scores, then finetune this base model with various sample sizes, a learning rate of 0.0001, a batch size of 16, and a weight decay of 3×10^{-6} on a single Nvidia A40 GPU. The Adam optimizer was employed with $\alpha = 0.9$, $\beta = 0.999$, and $\epsilon = 10^{-8}$. We exclusively use the Detoxify framework to evaluate the text generated by our model without the prompts.

D.3.1 A Semi-Synthetic Exercise Results

To study the effect of human-machine misalignment, we create a semi-synthetic data set of human and machine-generated toxicity scores. We use the Detoxify Hanu and Unitary team [2020] model as our "human" annotator, and a Detoxify model finetuned on a biased subsample of toxic instances as our "machine" annotator. For details about implementation, see Section F.1. Table 3 outlines ROC-AUC (Area Under ROC Curve) comparison between the original Detoxify model, and the finetuned model, which illustrates that the original Detoxify model outperforms the finetuned model in predicting

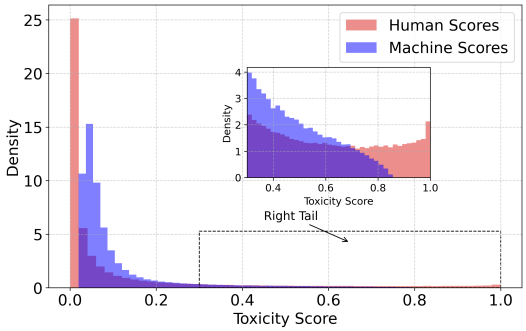

Figure 6: Distribution of toxicity scores assigned by humans (i.e., original Detoxify model) and machine (i.e., fine-tuned Detoxify model) for responses generated by the LLaMA2-7B model across all prompts. The main plot shows the density of scores, highlighting a long-tailed distribution in both human and machine evaluations. The zoomed-in plot illustrates that human scores have a heavier tail than machine scores.

toxicity. Fig. 6 illustrates the distribution of human and machine-assigned scores for responses generated by the LLaMA2-7B model across all sampled prompts. Both human and machine toxicity scores exhibit a long-tailed distribution in the real-world dataset. Due to the Detoxify model being fine-tuned with an emphasis on "severe toxicity" samples, its scores are predominantly concentrated in the lower range, highlighting a misalignment between human and machine assessments.

Table 3: Performance comparison of detoxify models.

Model	ROC-AUC
Detoxify (original)	0.97
Finetuned Detoxify ($\rho = 0.57$)	0.86

D.4 Choices of α

To select an appropriate value of α, we arrange $r_\lambda(x)$ in increasing order then rule out the top $q\%$ responses, and compute the desired distortion risk of the lower $(1-q)\%$. The calculated value is used as the target α. For details about implementation, see Section F.1. Table 4 outlines the rational α values under different settings of λ for different $q\%$ given that we are interested in studying CVaR.

Table 4: Choice of α with different $q\%$ under various settings of $\boldsymbol{\gamma}$ for CVaR.

$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$	$q = 1\%$	$q = 5\%$	$q = 10\%$	$q = 15\%$	$q = 20\%$
(3.07, 0.26, 32)	0.815	0.580	0.356	0.217	0.145
(3.18, 0.26, 32)	0.809	0.578	0.356	0.216	0.140
(3.31, 0.26, 32)	0.815	0.574	0.354	0.219	0.142
(3.48, 0.26, 32)	0.807	0.578	0.359	0.217	0.141
(3.76, 0.26, 32)	0.815	0.589	0.352	0.221	0.141

D.5 Deployment of the calibrated model

Recall that any choice of $\tilde{y}(x) \in \mathcal{C}_{\hat{\lambda}}(x)$ controls the risk $R_\psi(F_{\tilde{y}(x)})$. For a new prompt x, the most cost-effective approach is to sample candidate responses $y_1, y_2, \ldots$ from the underlying LLM $p(y \mid x)$ until the first time the machine disutility score is below $\hat{\lambda}$. Suppose each sample incurs a unit of computational cost. Given a prompt x, the sampling cost follows a geometric distribution with rate $\mathbb{P}(r_m(y) < \hat{\lambda} \mid x, \hat{\lambda})$. Therefore, on average, for a given prompt x, we expect the number of samples

Algorithm 2 Conformal distortion risk control

Input: machine-scoring model $r_m(\cdot)$, discrete subset of its range Λ, human-annotated scores $r(\cdot)$, set of prompts $\mathcal{X} = (x_i)_{i=1}^n$, target level α, tolerance level δ, weighting function ψ.
function DRC($\mathcal{X}, \Lambda, r_m(\cdot), r(\cdot), \alpha, \delta, \psi$)
for $x_i \in \mathcal{X}$ **do**
  $\mathcal{C} \leftarrow$ CANDIDATESET(x_i) {See Algorithm 1}
  for $\lambda \in \Lambda$ **do**
    $\mathcal{C}_\lambda(x_i) = \{\}$
    for $y_k \in \mathcal{C}(x_i)$ **do**
      if $r_m(y_k) < \lambda$ **then**
        $\mathcal{C}_\lambda(x_i) \leftarrow \mathcal{C}_\lambda(x_i) \cup \{y_k\}$
      end if
      $r_\lambda(x_i) \leftarrow \max\{r(y_j) : y_j \in \mathcal{C}_\lambda(x_i)\}$
    end for
  end for
end for
for $\lambda \in \Lambda$ **do**
  $(r_{\lambda,(1)}, \ldots, r_{\lambda,(n)}) \leftarrow$ SORT($r_\lambda(x_1), \ldots, r_\lambda(x_n)$)
  $\hat{R}_\psi(\lambda) \leftarrow \sum_{i=1}^n \left\{\psi\left(\frac{i}{n}\right) - \psi\left(\frac{i-1}{n}\right)\right\} r_{\lambda,(i)}$
  $\hat{\sigma}^2(\lambda) \leftarrow$ Eq. (5)
  $\hat{R}_\psi^+(\lambda) \leftarrow \hat{R}_\psi(\lambda) + z_{1-\delta} \cdot \frac{\hat{\sigma}(\lambda)}{\sqrt{n}}$ {UCB}
end for
$\hat{\lambda} \leftarrow \max\left\{\lambda \in \Lambda : \hat{R}_\psi^+(\lambda') \le \alpha,\ \forall \lambda' \le \lambda\right\}$
return $\hat{\lambda}$

needed to generate a response with disutility score less than $\hat{\lambda}$ is

$$N_x = \frac{1}{\mathbb{P}(r_m(y) < \hat{\lambda} \mid x, \hat{\lambda})}.$$

This demonstrates that the number of responses generated by our method, CDRC-L, is adaptive to the given prompt x, unlike best-of-N. In other words, the (unconditional) expected cost is

$$\mathrm{Cost}(\hat{\lambda}) = \mathbb{E}_x\left[\frac{1}{\mathbb{P}(r_m(y) < \hat{\lambda} \mid x, \hat{\lambda})}\right].$$

Suppose we have a hold-out set of prompts $\mathcal{D}'$ that is independent of the dataset $\mathcal{D}$ defined in Section 3.1, $\mathrm{Cost}(\hat{\lambda})$ can be estimated by

$$\frac{1}{|\mathcal{D}'|}\sum_{x \in \mathcal{D}'} \frac{1}{\widehat{\mathbb{P}}(r_m(y) < \hat{\lambda} \mid x, \hat{\lambda})}, \tag{7}$$

where the probability is estimated by the Monte-Carlo method.

E Algorithm details

Here we outline the Distortion Risk Control algorithm via L-statistics in Algorithm 2.

E.1 Alternative (conservative) approaches for distortion risk control

Another strategy to construct pointwise UCBs for $R_\psi(\lambda)$ is to replace all quantiles by their confidence envelopes. Following Snell et al. [2022], we consider two confidence envelopes based on the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality and Berk-Jones (BJ) statistics, respectively. For simplicity, we assume F_λ is continuous.

DKW inequality. The DKW inequality implies that, for any $\lambda \in \Lambda$ and $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{r\in\mathbb{R}} \mid \hat{F}_{n,\lambda}(r) - F_\lambda(r) \mid \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Letting $\epsilon_{n,\delta} = \sqrt{\log(2/\delta)/2n}$ and $r = q_p(\lambda)$ for $p \in [0,1]$, we obtain that, with probability at least $1-\delta$,

$$|\hat{F}_{n,\lambda}(q_p(\lambda)) - p| \leq \epsilon_{n,\delta}, \quad \forall p \in [0,1].$$

This implies an upper confidence envelope as $\hat{q}_p^+(\lambda) = r_{\lambda,(k_{n,p})}$, where $k_{n,p} = \lceil n(p+\epsilon_{n,\delta})\rceil$. Here, if $k_{n,p} > n$, we set $r_{\lambda,(k_{n,p})} = \lambda_{\max}$, the upper bound of Λ.

Berk-Jones Statistics. As pointed out by Snell et al. [2022] and Bates et al. [2023], the DKW inequality is overly conservative for p close to 0 and 1. The BJ statistic uses the fact that $F_r(\lambda_{\lambda,(i)}) \sim \mathrm{Beta}(i, n-i+1)$, where Beta denotes a Beta-distribution. Let $B_{i,n-i+1}$ denote the cumulative distribution function of $\mathrm{Beta}(i, n-i+1)$. The BJ statistic is then defined as

$$M_n^+ = \max_{1\leq i\leq n} G_{i,n-i+1}(F_r(r_{\lambda,(i)})).$$

Let s_δ be the δ-th quantile of M_n^+. Clearly, s_δ does not depend on F_r because $F_r(r_{\lambda,i}) \sim \mathrm{Unif}([0,1])$ when F_r is continuous. It can be computed in polynomial time Moscovich and Nadler [2017]. Let $s_i = G_{i,n-i+1}^{-1}(s_\delta)$. Then

$$\mathbb{P}(q_{s_i}(\lambda) \leq r_{\lambda,(i)},\ \forall i \leq n) \geq 1-\delta.$$

This yields an upper confidence envelope $\hat{q}_p^+(\lambda) = r_{\lambda,(i)}$ for any $p \in (s_{i-1}, s_i]$. For $p > s_n$, we set $\hat{q}_p^+(\lambda) = \lambda_{\max}$.

While both DKW and BJ approaches yield finite-sample valid UCBs for $R_\psi(\lambda)$, they are conservative because they do not target the specific choice of ψ. In fact, (3) implies that $\int_0^1 \hat{q}_p^+(\lambda)d\psi(p)$ is a uniform UCB across all weight functions, i.e.,

$$\mathbb{P}\left(R_\psi(\lambda) \leq \int_0^1 \hat{q}_p^+(\lambda)d\psi(p),\ \forall\psi\right) \geq 1-\delta.$$

Therefore, the actual coverage (i.e., probability that $R_\psi(\lambda)$ is less than or equal to the above UCB) is typically much higher than $1-\delta$. By contrast, the UCB given by the L-statistic is tailored to $R_\psi(\lambda)$ and the coverage converges to $1-\delta$ as $n \to \infty$.

F Experiments

In this section, we perform experiments to investigate the issue of human-machine misalignment by implementing our conformal distortion risk control method to mitigate toxicity of LLM-generated reseponses. This is a critical application, as toxic outputs may cause severely negative impacts on impressionable populations, moreover, propagate across wide audiences, leading to misinformation and harm.

F.1 Experimental setup

Datasets and models. We randomly draw 10K prompts from the REALTOXICITYPROMPTS dataset Gehman et al. [2020] . For each selected prompt x_i, we generate 40 responses $y_j(x_i)$ using the LLaMA2-7B model Touvron et al. [2023]. Given the initial responses, we apply the sequential algorithm described in Algorithm 1 to construct the candidate response sets $\mathcal{C}(x_i)$, ensuring the quality of the selected responses. Specifically, we use perplexity (PPL) to evaluate response quality, ROUGE-L to assess similarity between responses, and restrict the maximum set size to 32 as a stopping criterion. More details can be found in Appendix D.2.

Toxicity scores. To apply our method, we need a human toxicity score function $r(\cdot)$ and a machine toxicity score function $r_m(\cdot)$. An example of this disutility measure is illustrated in the Jigsaw Unintended Bias in Toxicity Classification cjadams et al. [2019] dataset, which provides toxicity labels from up to 10 human-annotators for each of the 2 million comments on a Civil Comments

platform. The total score $r(y(x))$ can be obtained by averaging over the ratings of the annotators. For general disutility metrics, $r(y(x))$ can also be defined as the negative reward estimated using the Bradley-Terry model, as done in RLHF.

Human-annotated data can be costly and time-consuming to acquire. To evaluate our method, we create a cheap semi-synthetic benchmark using an existing machine scoring model as the "human annotator," and a biased model as the "machine assessor." Specifically, we use the Detoxify model Hanu and Unitary team [2020] for $r(\cdot)$ and retrain the Detoxify model for $r_m(\cdot)$ on a biased subset of the Jigsaw Unintended Bias in Toxicity Classification dataset cjadams et al. [2019] that consists of the $c\%$ most and least toxic instances. The goal is to design $r_m(\cdot)$ with varying degrees of misalignment from $r(\cdot)$. This allows us to study the effect of misalignment on $\hat{\lambda}$ and hence the cost of tail risk control. In particular, we quantify the misalignment between human and machine scoring models by the Spearman correlation coefficient ρ between the scores across all candidate responses. In our experiment, we train three models for $r_m(\cdot)$ with $c\% \in \{15\%, 30\%, 70\%\}$. The Spearman correlation coefficients are $0.57, 0.68, 0.78$, respectively. More details about these models can be found in Appendix D.3.1.

Choices of parameters. We consider both CVaR_β and VaR_β control with $\beta \in \{0.5, 0.75, 0.9\}$. We fix the confidence parameter $1 - \delta = 0.95$. To determine a reasonable target level α, we compute the empirical CVaR_q on human scores of all candidate responses with $q \in \{1\%, 5\%, 10\%, 15\%, 20\%\}$; see Appendix D.4. This suggests a range of reasonable target levels. In particular, we consider $\alpha \in \{0.15, 0.2, 0.25, 0.3, 0.35\}$.

Evaluation. We randomly split the prompts, using $|\mathcal{D}| \in \{50, 100, 200, 1000, 6000\}$ to determine the optimal threshold $\hat{\lambda}$ and the remaining as a held-out test dataset. For each method, after selecting $\hat{\lambda}$, we deploy the calibrated model on the held-out dataset. We then evaluate the realized CVaR_β and VaR_β of human scores and estimate the sampling cost following (7). We apply all three versions of conformal distortion risk control based on L-statistics, DKW inequality, and BJ statistics. We refer to them as CDRC-L, CDRC-DKW, and CDRC-BJ, respectively, where CDRC stands for conformal distortion risk control. For CDRC-L and CDRC-DKW, we repeat for 15 times, and for CDRC-BJ, we repeat for 3 times due to computational complexity to compute s_δ.

Additional models. We conduct additional experiments using the LLaMA3.2-3B, and LLaMA3.1-8B models. Following the same implementation details as described, we obtain results that are highlighted in Appendix I.

G Additional experimental results for CVaR_β control

Realized risk and average cost analysis. Fig. 7, Fig. 8, and Fig. 9 show the realized CVaR_β of human scores and the average sampling cost with Spearman correlation, $\rho = 0.57$, on the held-out dataset as functions of α for $\beta \in \{0.5, 0.75, 0.9\}$ with $|\mathcal{D}| = 6000$. The panels in the first row shows that all methods control the risk at the target level and CDRC-L is least conservative, as discussed in Section 3.4. As a result, it incurs the smallest deployment cost among all three methods. Moreover, as β increases, the advantage of CDRC-L is more prominent. Although CDRC-BJ improves upon CDRC-DKW due to the tighter bounds for extreme quantiles, it still underperforms CRDC-L. On the other hand, the gap between the target and realized CVaR_β stays nearly constant for CDRC-L across different values of β, suggesting that L-statistics are adaptive to different choices of ψ.

Cost and misalignment. Next, we examine how the deployment cost varies with the misalignment between human and machine ratings. Fig. 10 demonstrates that, for all settings of β, as the Spearman correlation coefficient increases, the cost of generating a CVaR_β-controlled LLM response drops. This confirms our intuition that better-aligned machine ratings reduces the cost of calibration. Similar trends are observed for VaR_β control, as shown in Appendix H.

Comparison to best-of-N. Our work differs from inference-time alignment strategies, such as, best-of-N, a fixed-sample inference-time heuristic. In contrast, our method is an adaptive, risk-controlling strategy. In particular, best-of-N uses a fixed generation count N, meaning that it always samples N responses for each prompt. Conversely, our framework uses an adaptive number of samples, N_x,

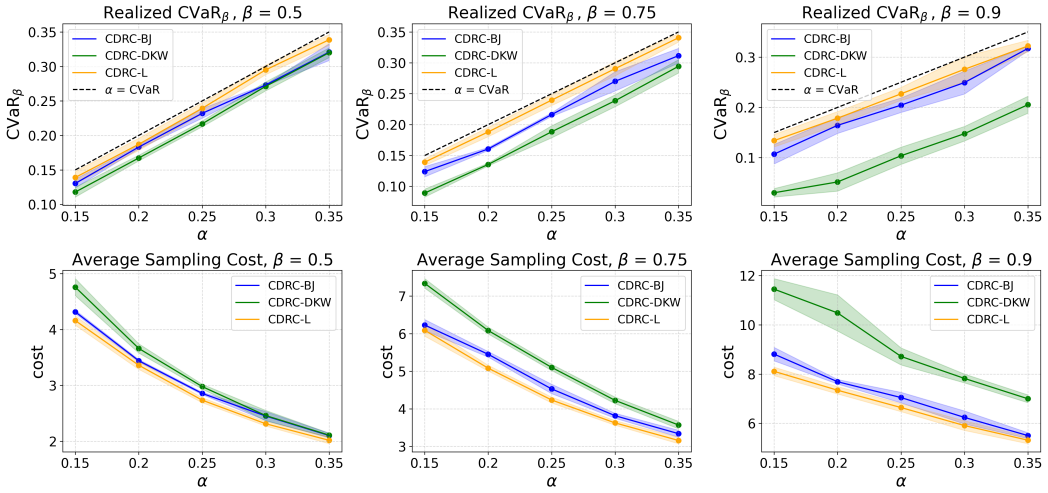

Figure 7: **LLaMA2-7B: Realized** CVaR_β **vs.** α **(row 1) and average sampling cost vs.** α **(row 2) for Spearman correlation between human and machine toxicity scores at** $\rho = 0.57$ **evaluated on held-out dataset with** $|\mathcal{D}| = 6000$. The confidence band is computed by taking the mean estimate plus/minus one standard error estimated from the results across independent experiments. Each subplot in the respective rows illustrates a different setting of $\beta \in \{0.5, 0.75, 0.9\}$. From the panels in the first row, we observe that our method, CDRC-L (orange), is an improvement to CDRC-DKW (green) and CDRC-BJ (blue), as it is able to achieve risk control (shown in the black dotted line) while being less conservative than both baseline methods. Evident from the panels in the second row, our method is more cost efficient and least conservative in generating a risk-controlled LLM response than DKW or BJ.

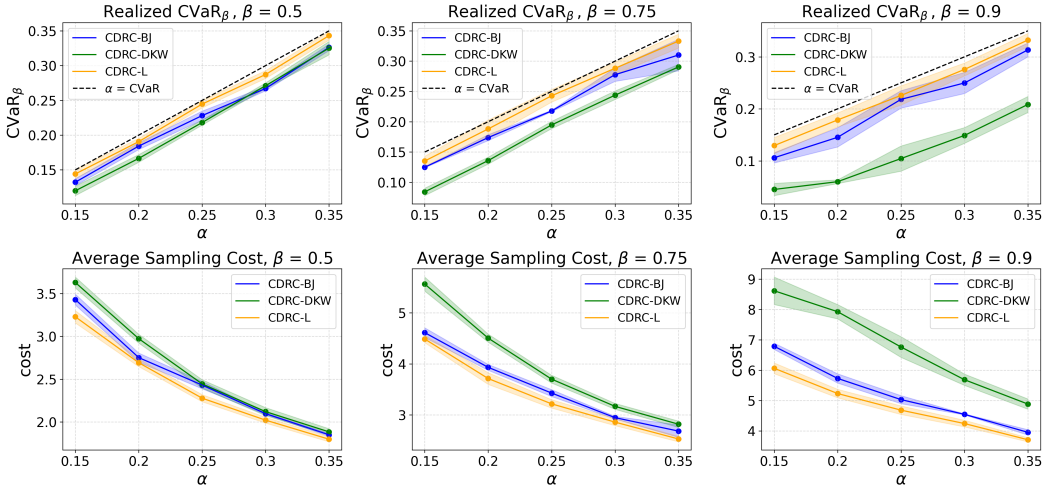

Figure 8: **Realized** CVaR_β **and average sampling cost on held-out dataset for** $\rho = 0.68$ with $|\mathcal{D}| = 6000$.

depending on the prompt and desired risk level. For example, when a prompt is non-toxic, best-of-N generates N responses, while our method may only require one, yielding a lower average inference cost. We illustrate this difference between the two methods in G. Furthermore, best-of-N implicitly assumes that the disutility function in the LLM framework is well-aligned with the human user. Not

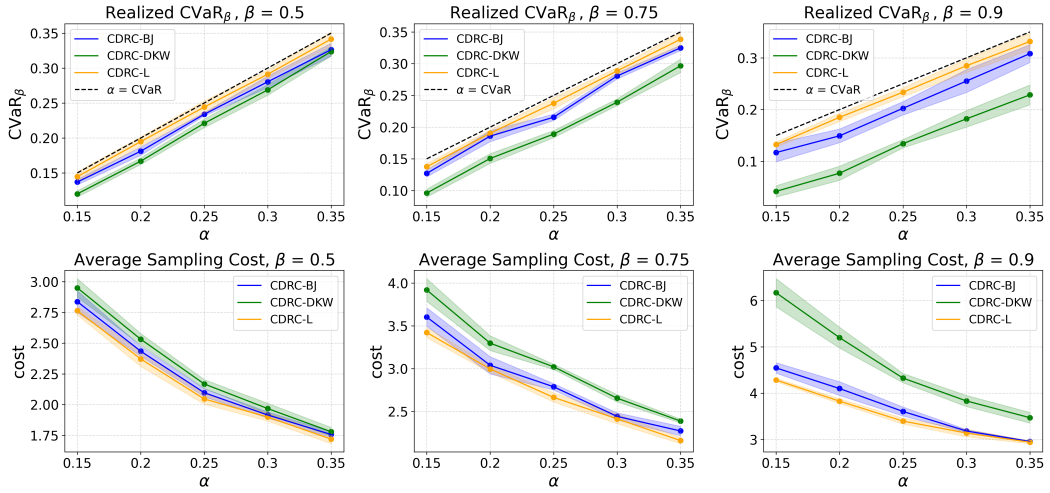

Figure 9: **Realized** CVaR_β **and average sampling cost on held-out dataset for** $\rho = 0.78$ with $|\mathcal{D}| = 6000$.

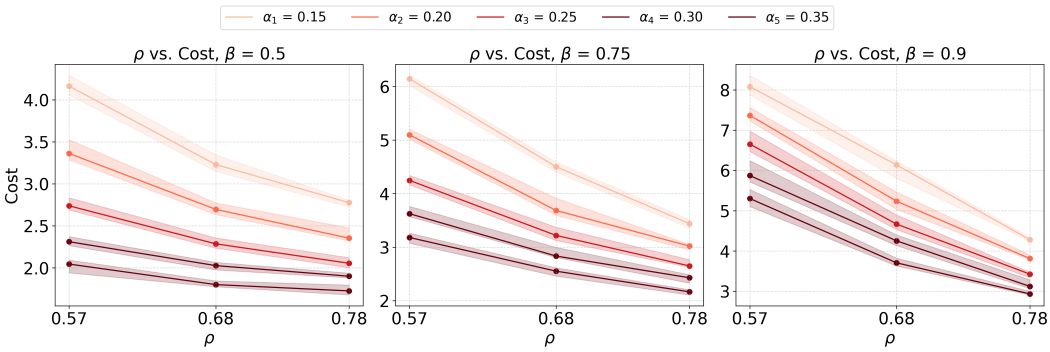

Figure 10: **LLaMA2-7B: Spearman correlation vs. average sampling cost for** CVaR_β **with** $|\mathcal{D}| = 6000$.

only does our method not make such assumptions, but we also extend the problem of alignment by introducing and tuning a parameter that effectively controls the human distuility of an LLM generated response.

We compared the two methods under a controlled setup: setting $\beta = 0.5$, using $N \in 3, 5$ for best-of-N, and calibration set size $n = 1000$. Following the procedure in F.1, we evaluated the realized CVaR_β of human-annotated toxicity scores using the response with the lowest machine score from each group of N samples:

Best-of-N	Mean $\mathrm{CVaR}_\beta \pm$ standard deviation
Best-of-3	0.2427 ± 0.0068
Best-of-5	0.2015 ± 0.007

Table 5: Beta CVaR_β Human Scores with standard deviation across trials.

The results in Table 5 demonstrate that our method achieves comparable CVaR_β control with lower average inference cost than best-of-N (see Fig. 7). In particular, on average, for $\beta = 0.5$, best-of-3 controls CVaR_β at approximately level $\alpha = 0.25$, and CDRC-L can achieve the same level of control by generating less than 3 LLM responses. Furthermore, on average, for $\beta = 0.5$, best-of-5 controls

CVaR_β at approximately level $\alpha = 0.20$, while CDRC-L can achieve the same level of control by generating less than 3.5 responses.

H Distortion Risk Control for VaR_β

H.1 Asymptotics of empirical quantiles

Take $\psi(q) = \delta_\beta(q)$, we obtain $\mathrm{VaR}_\beta(\lambda) = F_\lambda^{-1}(\beta)$. Theorem B.1 implies that the empirical β-th quantile is asymptotically normal with variance

$$\sigma^2_{\mathrm{VaR}_\beta}(\lambda) = \frac{\beta(1-\beta)}{f^2_\lambda(F^{-1}_\lambda(\beta))}.$$

Empirically, we use the standard Bootstrap procedure to estimate $\sigma^2_{\mathrm{VaR}_\beta}(\lambda)$ with 1000 samples.

H.2 Additional experimental results for VaR_β

We evaluate the performance, i.e., average sampling cost, and realized VaR_β on a held-out dataset with $|\mathcal{D}| = 6000$, of our framework CDRC-L, against baseline methods CDRC-DKW and CDRC-BJ. See Fig. 12 for $\rho = 0.68$, and Fig. 13 for $\rho = 0.78$. Similar to what we observe with CVaR_β, it is evident from the plots that all methods control the risk at the specified level α since they fall below the black dotted line (where $\alpha = \mathrm{VaR}_\beta$) within the margin of error. Again, our method, CDRC-L, is consistently the least conservative of all methods across different values of β, and ρ. As a result, we see that our method is consistently the least costly across all settings of β, and ρ. Fig. 14 illustrates that for all settings of β, as the Spearman correlation coefficient increases, the cost of generating a VaR_β-controlled LLM response decreases.

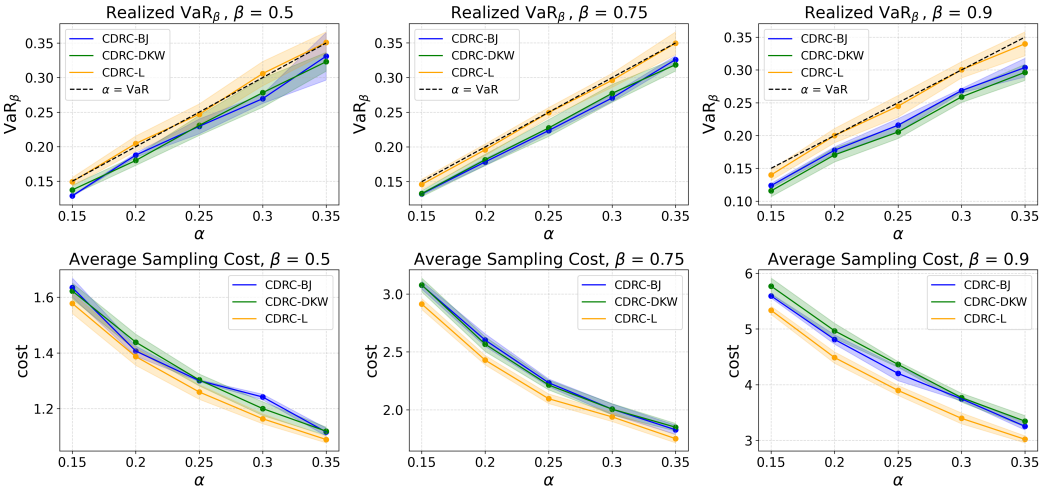

Figure 11: **Realized VaR_β vs. α (row 1) and average sampling cost vs. α (row 2), and for Spearman correlation between human and machine toxicity scores at $\rho = 0.57$ evaluated on held-out dataset with $|\mathcal{D}| = 6000$.** The confidence band is computed by taking the mean estimate plus/minus one standard error estimated from the results across the independent experiments. Each panel in the respective rows illustrates a different setting of $\beta \in \{0.5, 0.75, 0.9\}$. CDRC-L (orange), is an improvement to CDRC-DKW (green) and CDRC-BJ (blue), as it is able to achieve risk control within the margin of error (shown in the black dotted line) while being more cost efficient and less conservative than both baseline methods.

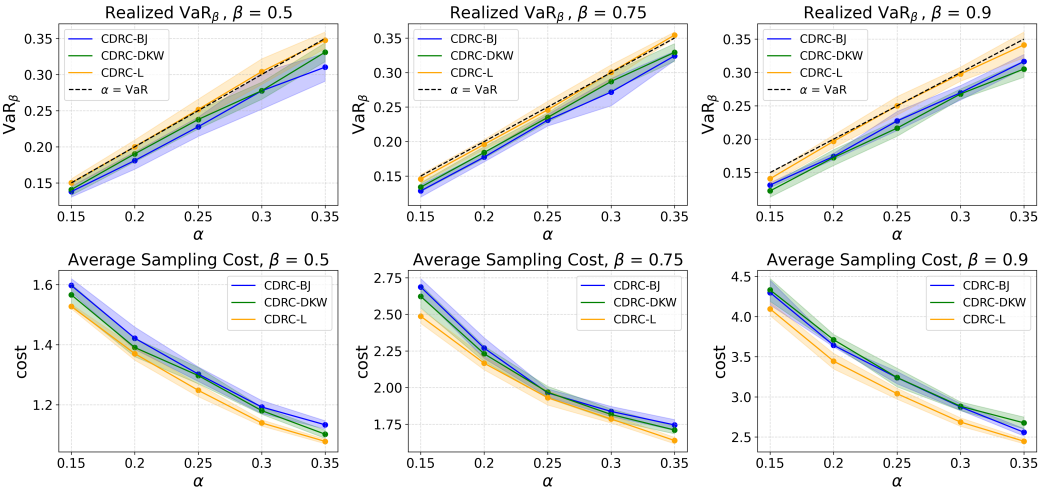

Figure 12: **Average cost and realized** VaR_β **on held-out dataset for** $\rho = 0.68$ **with** $|\mathcal{D}| = 6000$**.**

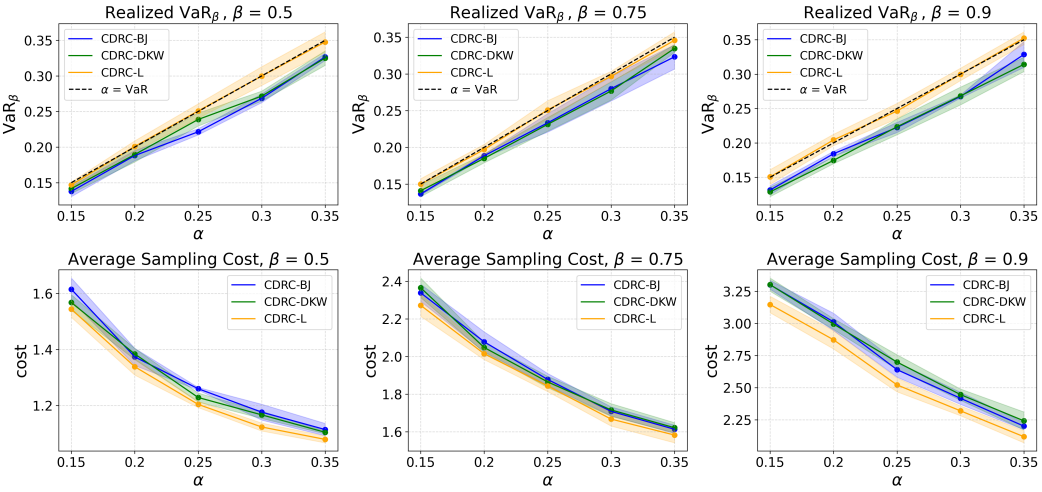

Figure 13: **Average cost and realized** VaR_β **on held-out dataset for** $\rho = 0.78$ **with** $|\mathcal{D}| = 6000$.

Similar to the results for CVaR_β, from Fig. 15, it is evident that CDRC-L is consistently the least conservative among all methods, while still maintaining the guaranteed VaR_β control within the margin of error. Furthermore, from Fig. 16, we observe that CDRC-L is consistently the most cost effective among all methods. In particular, it is significantly less expensive in comparison to other methods in settings with small calibration set sizes.

I Additional model results

I.1 LLaMA3.1-8B

To evaluate our tail risk control framework using the LlaMA3.1-8B model, we choose $\alpha \in \{0.125, 0.15, 0.2, 0.25, 0.3\}$ for $\beta \in \{0.5, 0.75\}$, and $\alpha \in \{0.2, 0.25, 0.3\}$ for $\beta = 0.9$ for testing CVaR_β control. In addition, we choose $\alpha \in \{0.05, 0.075, 0.1\}$ for $\beta = 0.5$, and

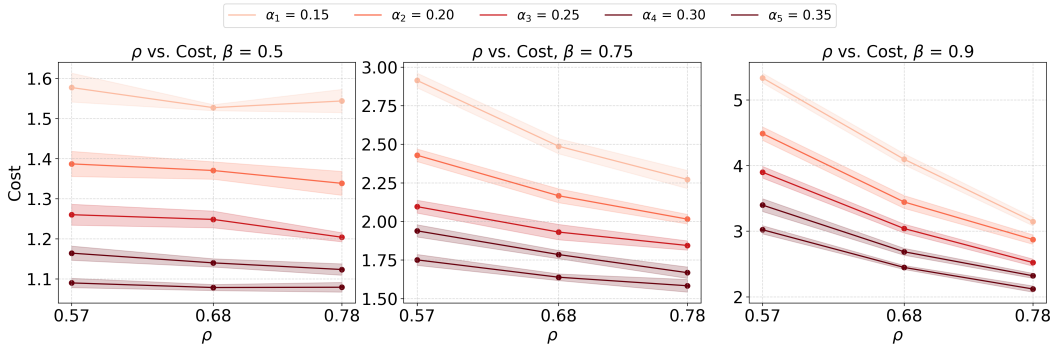

Figure 14: **Spearman correlation vs. average sampling cost for** VaR_β **with** $|\mathcal{D}| = 6000$.

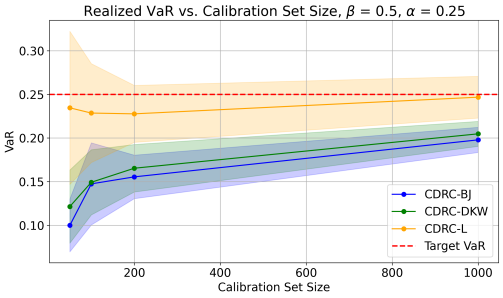

Figure 15: **Realized** VaR_β **vs. calibration set size** for $\beta = 0.5$, $\alpha = 0.25$, $\rho = 0.57$.

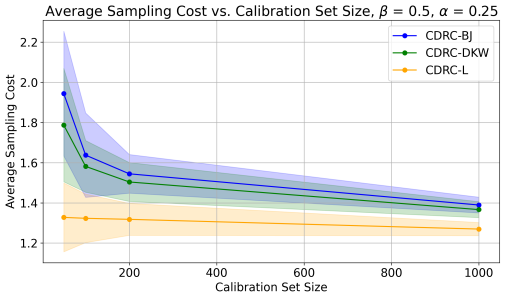

Figure 16: **Average sampling cost for** VaR_β **control vs. calibration set size** for $\beta = 0.5$, $\alpha = 0.25$, $\rho = 0.57$.

$\alpha \in \{0.1, 0.15, 0.2, 0.25, 0.3\}$ for $\beta \in \{0.75, 0.9\}$ for testing VaR_β control. Using $\rho = 0.57$, we illustrate the results for CVaR_β and VaR_β control in 17 and 18, respectively for $|\mathcal{D}| = 6000$.

We examine the performance, i.e. the realized CVaR_β and VaR_β, and average sampling cost of our framework, CDRC-L, on a held-out dataset against baseline methods CDRC-DKW and CDRC-BJ. Similar to what we observe with the LlaMA2-7B model, it is evident from the plots that all methods control the risk at the specified level α since they fall below the black dotted line (where $\alpha = \mathrm{CVaR}_\beta$ or VaR_β) within the margin of error. Meanwhile, our method, CDRC-L, is consistently the least conservative and hence most cost efficient of all methods across different values of β.

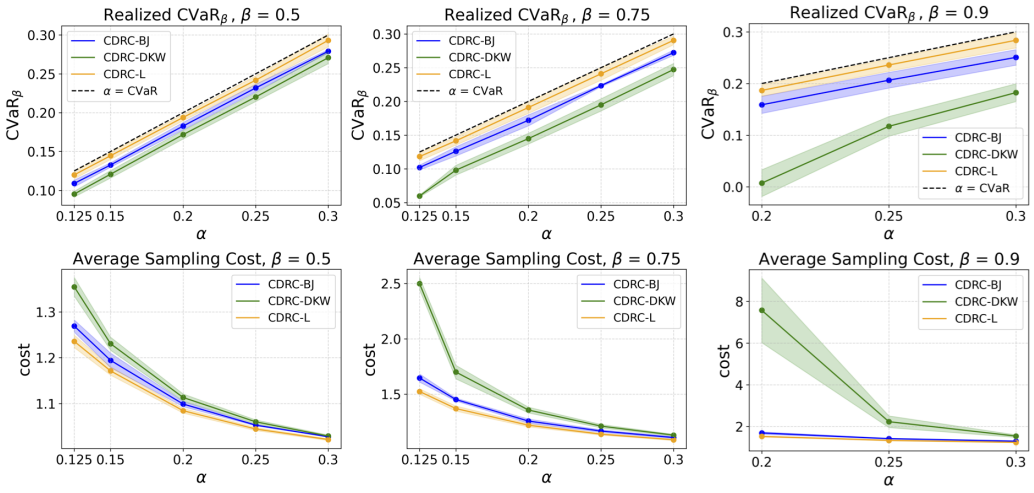

Figure 17: **LLaMA3.1-8B:Realized** CVaR_β **vs.** α **(row 1), and average sampling cost vs.** α **(row 2) for Spearman correlation between human and machine toxicity scores at** $\rho = 0.57$ **evaluated on held-out dataset with** $|\mathcal{D}| = 6000$. Each panel in the respective rows illustrates a different setting of $\beta \in \{0.5, 0.75, 0.9\}$. CDRC-L (orange), is an improvement to CDRC-DKW (green) and CDRC-BJ (blue), as it is able to achieve risk control within the margin of error (shown in the black dotted line) while being more cost efficient and less conservative than both baseline methods.

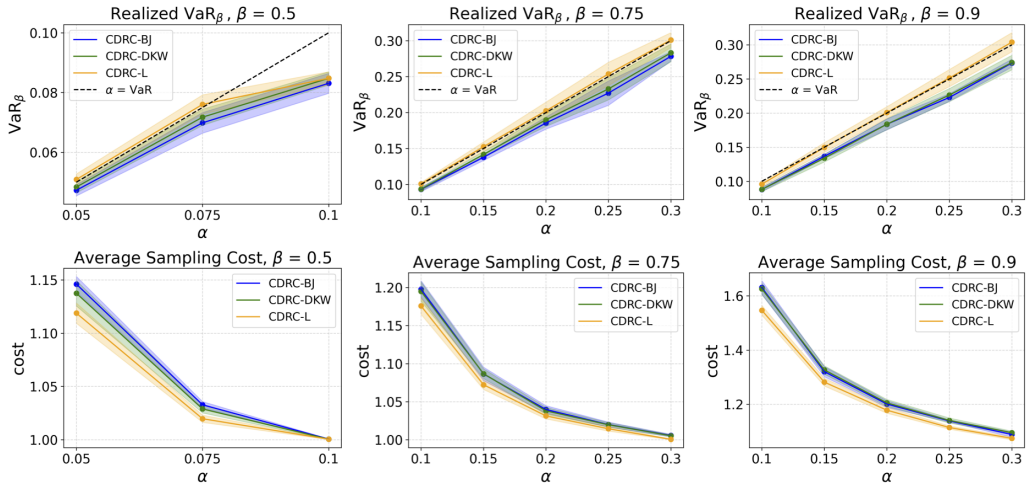

Figure 18: **LLaMA3.1-8B: Realized** VaR_β **vs.** α **(row 1), and average sampling cost vs.** α **(row 2), and for Spearman correlation between human and machine toxicity scores at** $\rho = 0.57$ **evaluated on held-out dataset with** $|\mathcal{D}| = 6000$. Each panel in the respective rows illustrates a different setting of $\beta \in \{0.5, 0.75, 0.9\}$. CDRC-L (orange), is an improvement to CDRC-DKW (green) and CDRC-BJ (blue), as it is able to achieve risk control within the margin of error (shown in the black dotted line) while being more cost efficient and less conservative than both baseline methods.

I.2 Llama3.2-3B

To evaluate our tail risk control framework using the LlaMA3.2-3B model, we choose $\alpha \in \{0.15, 0.2, 0.25, 0.3, 0.35\}$ for $\beta \in \{0.5, 0.75\}$, and $\alpha \in \{0.2, 0.25, 0.3\}$ for $\beta = 0.9$ for testing CVaR_β control. In addition, we choose $\alpha \in \{0.15, 0.2, 0.25, 0.3, 0.35\}$ for $\beta \in \{0.5, 0.75, 0.9\}$

for testing VaR_β control. Using $\rho = 0.57$, we illustrate the results for CVaR_β and VaR_β control in 19 and 20, respectively for $|\mathcal{D}| = 6000$.

Again, we examine the performance, i.e. the realized CVaR_β and VaR_β, and average sampling cost of our framework, CDRC-L, on a held-out dataset against baseline methods CDRC-DKW and CDRC-BJ. Similar to what we observe with the LlaMA2-7B and LlaMA3.1-8B model, it is evident from the plots that all methods control the risk at the specified level α since they fall below the black dotted line (where $\alpha = \mathrm{CVaR}_\beta$ or VaR_β) within the margin of error. Meanwhile, our method, CDRC-L, is consistently the least conservative and therefore, the most least costly of all methods across different values of β.

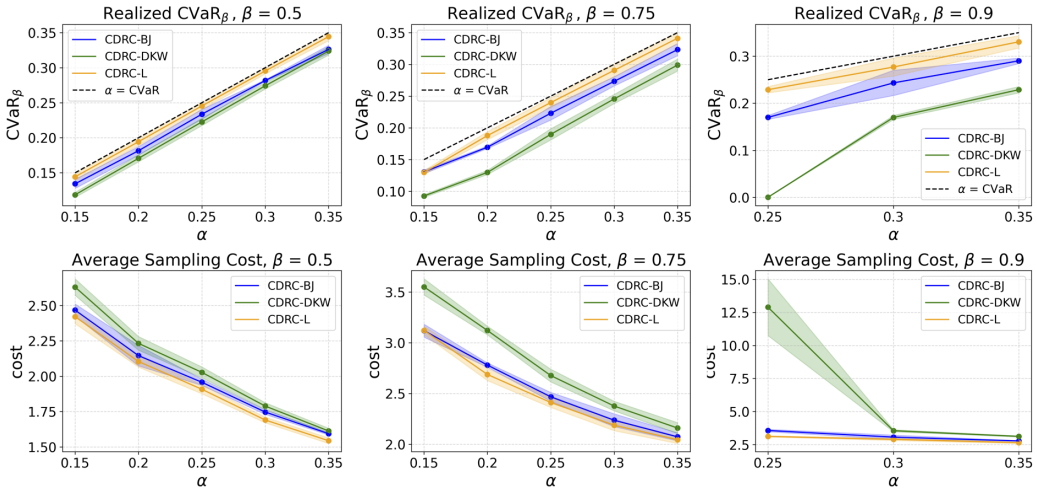

Figure 19: **Llama3.2-3B: Realized CVaR_β vs. α (row 1), and average sampling cost vs. α (row 2) for Spearman correlation between human and machine toxicity scores at $\rho = 0.57$ evaluated on held-out dataset with $|\mathcal{D}| = 6000$.** Each panel in the respective rows illustrates a different setting of $\beta \in \{0.5, 0.75, 0.9\}$. CDRC-L (orange), is an improvement to CDRC-DKW (green) and CDRC-BJ (blue), as it is able to achieve risk control within the margin of error (shown in the black dotted line) while being more cost efficient and less conservative than both baseline methods.

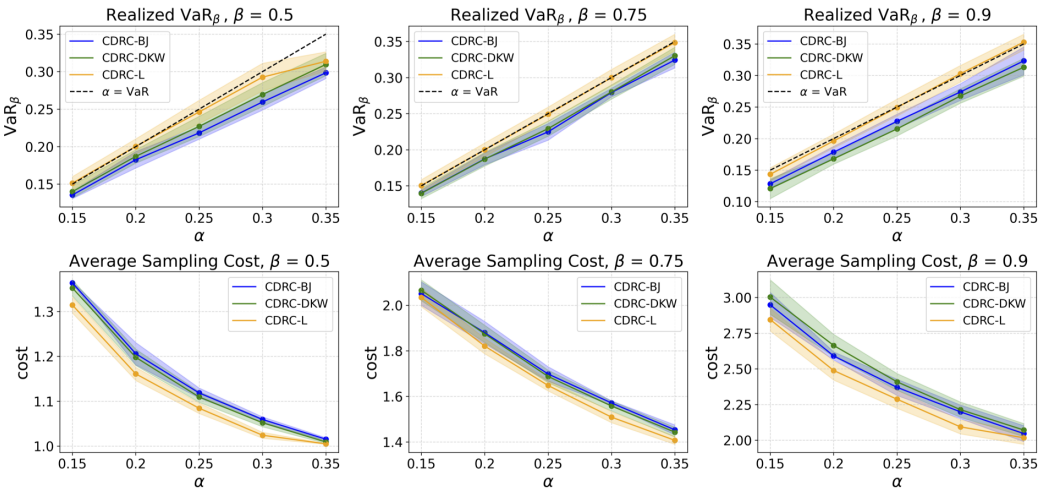

Figure 20: **Llama3.2-3B: Realized** VaR_β **vs.** α **(row 1), and average sampling cost vs.** α **(row 2) for Spearman correlation between human and machine toxicity scores at** $\rho = 0.57$ **evaluated on held-out dataset with** $|\mathcal{D}| = 6000$. Each panel in the respective rows illustrates a different setting of $\beta \in \{0.5, 0.75, 0.9\}$. CDRC-L (orange), is an improvement to CDRC-DKW (green) and CDRC-BJ (blue), as it is able to achieve risk control within the margin of error (shown in the black dotted line) while being more cost efficient and less conservative than both baseline methods.