Risk Aware Benchmarking of Large Language Models

Apoorva Nitsure [1] **Youssef Mroueh** [1] **Mattia Rigotti** [1] **Kristjan Greenewald** [1 2] **Brian Belgodere** [1]
Mikhail Yurochkin [1 2] **Jiri Navratil** [1] **Igor Melnyk** [1] **Jarret Ross** [1]

Abstract

We propose a distributional framework for benchmarking socio-technical risks of foundation models with quantified statistical significance. Our approach hinges on a new statistical relative testing based on first and second order stochastic dominance of real random variables. We show that the second order statistics in this test are linked to mean-risk models commonly used in econometrics and mathematical finance to balance risk and utility when choosing between alternatives. Using this framework, we formally develop a risk-aware approach for foundation model selection given guardrails quantified by specified metrics. Inspired by portfolio optimization and selection theory in mathematical finance, we define a *metrics portfolio* for each model as a means to aggregate a collection of metrics, and perform model selection based on the stochastic dominance of these portfolios. The statistical significance of our tests is backed theoretically by an asymptotic analysis via central limit theorems instantiated in practice via a bootstrap variance estimate. We use our framework to compare various large language models regarding risks related to drifting from instructions and outputting toxic content.

1. Introduction

Foundation models such as large language models (LLMs) have shown remarkable capabilities redefining the field of artificial intelligence. At the same time, they present pressing and challenging socio-technical risks regarding the trustworthiness of their outputs and their alignment with human values and ethics (Bommasani et al., 2021). Evaluating LLMs is therefore a multi-dimensional problem, where those risks are benchmarked across diverse tasks and domains (Chang et al., 2023).

In order to quantify these risks, (Liang et al., 2022; Wang et al., 2023; Huang et al., 2023; Sun et al., 2024) proposed benchmarks of automatic metrics for probing the trustworthiness of LLMs. These metrics include accuracy, robustness, fairness, toxicity of the outputs, etc. Human evaluation benchmarks can be even more nuanced, and are often employed when tasks surpass the scope of standard metrics. Notable benchmarks based on human and automatic evaluations include, among others, Chatbot Arena (Zheng et al., 2023), HELM (Bommasani et al., 2023), MosaicML's Eval, Open LLM Leaderboard (Wolf, 2023), and BIG-bench (Srivastava et al., 2022), each catering to specific evaluation areas such as chatbot performance, knowledge assessment, and domain-specific challenges. Traditional metrics, however, sometimes do not correlate well with human judgments. Aiming for a better alignment with human judgments, some approaches utilize ChatGPT/GPT-4 for natural language generation evaluations (Liu et al., 2023; Zhang et al., 2023; Hada et al., 2023).

A comprehensive evaluation of LLMs requires addressing the following critical considerations:

1. ***Interpretability.*** Evaluation of foundation models is multi-dimensional in nature and multiple metrics benchmark the models on different socio-technical dimensions that probe the trustworthiness of their outputs and their adherence to shared values and ethics. *It is critical to establish an aggregate-level measure to facilitate the interpretation and effective communication of the evaluation results.*

2. ***Risk Aware Benchmarking.*** In natural language (and other) applications, metrics quantify important guardrails such as model's toxicity, safety, or robustness. Therefore, a comprehensive evaluation framework must incorporate a risk aware benchmarking. This entails ranking models based on the assessment of failure modes and tail statistics[1], providing a nuanced understanding of potential pitfalls.

[1]IBM Research [2]MIT-IBM Watson AI Lab. Correspondence to: Apoorva Nitsure <Apoorva.Nitsure@ibm.com>, Youssef Mroueh <mroueh@us.ibm.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]I.e. understanding and quantifying low-probability high-risk events.

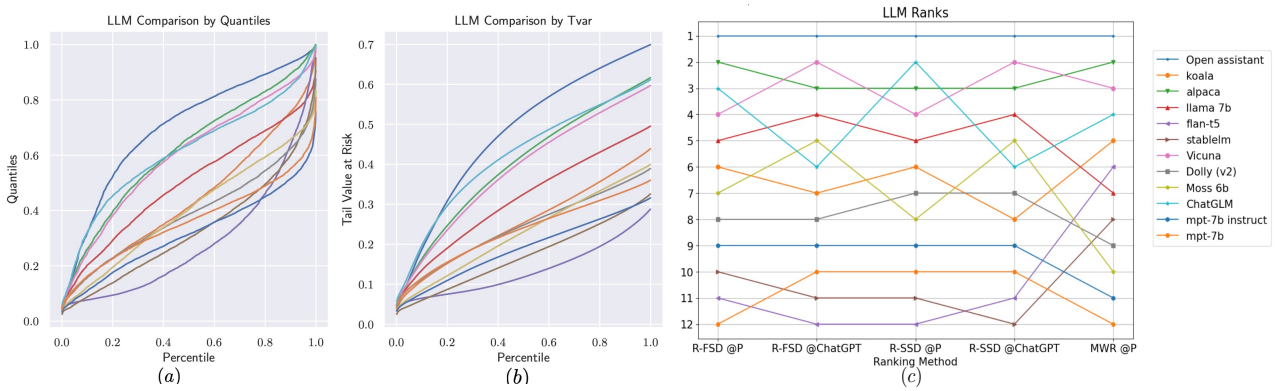


Figure 1: (a) Quantiles, (b) Tail Value at Risk (TVAR), of Metrics portfolio of an LLM, showing that TVAR (second-order stochastic dominance) more clearly ranks the models than the quantiles alone (first-order stochastic dominance). (c) Ranking of models using Relative First and Second Stochastic Dominance of Portfolios (R-FSD, R-SSD @P) versus ranking of models using Relative First and Second Stochastic Dominance of chatGPT evaluation scores and ranking by Mean Win Rate (MWR) on the metrics portfolio. The portfolio in this plot uses an independent copula aggregation. Note that (1) the metrics portfolio successfully approximates the chatGPT evaluation, since the @P rankings largely agree with the @chatGPT rankings; (2) the R-SSD rankings outperform MWR baseline.

3. ***Statistical Significance.*** Evaluating machine learning models is intimately connected to statistical significance testing (SST), although this framework is still underutilized: (Dror et al., 2018) reports almost 50% of ACL papers miss SST indicators. With the ever increasing parametric complexity of LLMs, obtaining a reliable SST in evaluating foundation models becomes ever more urgent.

We propose in this paper an evaluation framework that offers a principled solution and an efficient implementation that addresses each of these challenges. Our main contributions are:

1. ***Interpretable Metrics-Portfolio (Section 4).*** Drawing inspiration from econometrics and mathematical finance, we define a metrics-portfolio for aggregating metrics. This portfolio uses the notion of *copula* to normalize and aggregate metrics, yielding a single interpretable number assessing each output of a LLM. A higher value of the portfolio is preferable. We illustrate in Figure 1 panels (a) and (b) summary statistics of the metrics portfolio aggregating a total of 8 automatic metrics computed using $5K$ samples from the Mix-instruct dataset (Jiang et al., 2023). In panel (c) we show that model ranking based on our metrics-portfolio aligns with human evaluation proxies such as chatGPT (Please refer to Appendix B for details of how chatGPT score is computed).

2. ***Risk Aware Benchmarking via Second Order Stochastic Dominance (Section 2).*** Stochastic orders define partial orders on random variables and play a vital role in econometrics and mathematical finance for comparing and selecting portfolios. We propose using stochastic order to select LLMs based on their metrics-portfolios. A portfolio dominates in the First Order Stochastic Dominance (FSD) if it has higher quantiles for all percentiles. However, in Figure 1 (Panel (a)), the quantiles of the metrics-portfolio of an LLM don't provide a clear ordering. Instead, we propose the use of Second Stochastic Dominance (SSD), where a portfolio dominates if it has higher Tail Values at Risk (TVAR) for all percentiles (also known as Conditional Value at Risk). TVAR, illustrated in Figure 1 (Panel (b)), represents normalized integrated quantiles, assessing the risks of low values in the portfolio. Small TVAR corresponds to fat left tails in the distribution of the portfolio, identifying risky LLMs as those with the lowest TVAR. For example, Flan-T5 emerges as the riskiest model in our running example.

3. ***Statistical Significance via Dominance Tests. (Section 3)*** Armed with these notions of stochastic dominance, we define statistics that benchmark the *relative* dominance of a model's portfolio on another (R-FSD and R-SSD in Panel (c) in Figure 1). We subject these statistics to an asymptotic analysis, proving central limit theorems that provide the foundation for hypothesis testing with false discovery rate control. We then perform stochastic dominance hypothesis testings between all pairs of models. Having adjusted the confidence level of these tests, we aggregate these pairwise rankings to a single rank via rank aggregation techniques such as the Borda Algorithm (de Borda, 1781). The resulting ranks, depicted in Panel (c) of Figure 1, highlight that the portfolio of automatic metrics (@P) leads

to a similar ranking to chatGPT score (@chatGPT) for both first and second stochastic order. To underscore the importance of risk aware benchmarking, we present the ranking of the metrics-portfolio produced by the ubiquitous Min Win Rate (MWR) used in LLM benchmarks (Liang et al., 2022)(last column in Panel (c)). Flan-T5 ranks close to last with all other orders, but ranks 6 with MWR. This highlights that the ubiquituous MWR used in LLM benchmarks is risky for ranking LLMs as it does not take into account failure modes of the model, and we caution practitioners of its pitfalls.

2. Stochastic Dominance

We first review notions of stochastic dominance and their relation to downside risk measures and risk averse preference modeling. We use the notation of the seminal paper of (Ogryczak & Ruszczynski, 2002), and assume that the random variables are standardized so that larger outcomes are preferable. Throughout this Section, the reader can think of the random variable X as a metric evaluating the performance of model A on a specific test set. Likewise, Y represents the evaluation of model B. We defer the definition of metrics portfolio to Section 4. In a multi-metric evaluation, as explained in the introduction, X and Y represent portfolios of evaluations of model A and B respectively.

2.1. First and Second order Dominance and Mean-Risk Models

First Order Stochastic Dominance The First-order Stochastic Dominance (FSD) between real-valued random variables uses the right-continuous cumulative distribution (CDF) as a performance function. Specifically, for a real random variable X, define the first performance function $F_X^{(1)} : \mathbb{R} \to [0,1]$ as the CDF: $F_X^{(1)}(\eta) = \mathbb{P}(X \leq \eta), \forall \eta \in \mathbb{R}$. The FSD of X on Y is defined as follows:

$$X \underset{\text{FSD}}{\succcurlyeq} Y \iff F_X^{(1)}(\eta) \leq F_Y^{(1)}(\eta), \forall \eta \in \mathbb{R}, \quad (1)$$

this intuitively means that for all outcomes η, the probability of observing smaller outcomes than η is lower for X than Y. An equivalent definition can be expressed using the quantile $F_X^{(-1)}$ (See e.g (Ogryczak & Ruszczynski, 2002)):

$$X \underset{\text{FSD}}{\succcurlyeq} Y \iff F_X^{(-1)}(p) \geq F_Y^{(-1)}(p), \forall p \in (0,1], \quad (2)$$

where $F_X^{(-1)} : [0,1] \to \overline{\mathbb{R}}$ is the left-continuous inverse of $F_X^{(1)}$: $F_X^{(-1)}(p) = \inf\{\eta : F_X^{(1)}(\eta) \geq p\}$ for $p \in (0,1]$. We focus on this definition as it is more computationally and notationally friendly since the quantile function is always supported on $[0,1]$.

Second Order Stochastic Dominance The Second-order Stochastic Dominance (SSD) is defined via the second performance function $F_X^{(2)} : \mathbb{R} \to [0,1]$ that measures the area under the CDF: $F_X^{(2)}(\eta) = \int_{-\infty}^{\eta} F_X^{(1)}(x)dx$, for $x \in \mathbb{R}$, yielding:

$$X \underset{\text{SSD}}{\succcurlyeq} Y \iff F_X^{(2)}(\eta) \leq F_Y^{(2)}(\eta), \forall \eta \in \mathbb{R}. \quad (3)$$

Note that FSD implies SSD, hence SSD is a finer notion of dominance. While FSD implies that X is preferred to Y by any utility-maximizing agent preferring larger outcomes[2], (Ogryczak & Ruszczynski, 2002) showed that SSD implies that X is preferred to Y by any *risk-averse* agent preferring larger outcomes.[3] Similarly to FSD, SSD can be measured with quantile functions via introducing the second quantile function also known as *integrated quantiles* $F_X^{(-2)} : (0,1] \to \overline{\mathbb{R}}$

$$F_X^{(-2)}(p) = \int_0^p F_X^{(-1)}(t)dt, \text{ for } t \in (0,1]. \quad (4)$$

Similarly to the FSD case, an equivalent more computationally friendly definition can be expressed in terms of the second quantile function (a proof of this equivalence can be found in Theorem 3.2 in (Ogryczak & Ruszczynski, 2002)):

$$X \underset{\text{SSD}}{\succcurlyeq} Y \iff F_X^{(-2)}(p) \geq F_Y^{(-2)}(p), \forall p \in (0,1]. \quad (5)$$

This equivalence is not straightforward and is due to Fenchel duality between $F^{(2)}$ and $F^{(-2)}$. Using $p = 1$ we see that SSD implies $\mu_X \geq \mu_Y$, where μ_X and μ_Y are means of X and Y.

Mean – Risk Models (MRM) As noted earlier SSD is linked to risk aware benchmarking via the second performance function $F^{(2)}(.)$ measuring expected shortfall, and the negative second quantile function $-F^{(-2)}(p)$ that is an assessment of expected losses given outcomes lower than the p-quantile.

Definition 2.1 (Mean – Risk Models)**.** A mean – risk model of a random variable X consists of the pair (μ_X, r_X), where μ_X is the mean of X, and r_X is a functional that measures the risk of the random outcome X.

The consistency of a mean – risk model with SSD is defined as follows:

[2] I.e. having an increasing utility function.

[3] I.e. having an increasing and *concave* utility function.

Name	Risk Measure	$\alpha-$ consistency with SSD
Standard deviation	$\sigma_X = \sqrt{\mathbb{E}(X-\mu_X)^2}$	not consistent
Absolute semi deviation	$\delta_X = \mathbb{E}(\mu_X - X)_+$	$1-$ consistent
Negative Tail Value at Risk	$-\text{TVAR}_X(p) = -\frac{F^{(-2)}(p)}{p}$	$1-$ consistent for all $p \in (0,1]$
Mean absolute deviation from a quantile	$h_X(p) = \mu_x - \frac{F_X^{(-2)}(p)}{p}$	$1-$ consistent for all $p \in (0,1]$
Gini Tail	$\Gamma_X = 2\int_0^1 (\mu_X p - F_X^{(-2)}(p))dp$	$1-$ consistent

Table 1: Risk models and their $\alpha-$consistency with SSD.

Definition 2.2 (SSD consistency of Mean – Risk Models)**.** A mean – risk model (μ_X, r_X) is $\alpha-$consistent with SSD, if for $\alpha > 0$ the following is true:

$$X \underset{\text{SSD}}{\succcurlyeq} Y \implies \mu_X - \alpha r_x \geq \mu_Y - \alpha r_Y. \tag{6}$$

The ubiquitous mean – risk model in machine learning is (μ_X, σ_X), where σ_X is the standard deviation. Unfortunately this model is not consistent with the SSD and has several limitations as it implies Gaussianity of the outcomes or a quadratic utility function. We give in Table 1 risk measurements and their $\alpha-$consistency (proofs in (Ogryczak & Ruszczynski, 2002)). Note that in contrast FSD is only consistent with the Mean-VaR risk model (Mean-Value at Risk) for all $p \in [0,1]$. VaR does not provide a refined tail assessment.

2.2. Relaxations of Stochastic Dominance

Recalling the definitions of FSD and SSD in Equations (2) and (5), in the finite-sample regime it is hard to test for these relations as one needs to show the infinite-sample quantile or second quantile properties hold uniformly over all $p \in (0,1]$. This difficulty motivated the relaxation of stochastic dominance to an almost stochastic dominance pioneered by (Leshno & Levy, 2002). These relaxations were revisited for the first order by (Alvarez-Esteban et al., 2014) who later proposed an optimal transportation approach to assess almost first stochastic order (Del Barrio et al., 2018).

Almost FSD (ε-FSD) Following (Leshno & Levy, 2002), (Del Barrio et al., 2018) relaxed FSD (Equation (2)) via the violation ratio of FSD. $X \underset{\varepsilon-\text{FSD}}{\succcurlyeq} Y$ if and only if:

$$\varepsilon_{\mathsf{W}_2}(F_X, F_Y) = \frac{\int_0^1 (F_Y^{(-1)}(t) - F_X^{(-1)}(t))_+^2 dt}{\mathsf{W}_2^2(F_X, F_Y)} \leq \varepsilon, \tag{7}$$

where W_2 is the Wasserstein -2 distance between F_X and F_Y.This ratio corresponds to a measure of the "area" of violation of the FSD dominance of X on Y. Note that $0 \leq \varepsilon_{\mathsf{W}_2}(F_X, F_Y) \leq 1$, with value 0 if $X \underset{\text{FSD}}{\succ} Y$ and 1 if $Y \underset{\text{FSD}}{\succ} X$. For $\varepsilon \in (0, \frac{1}{2}]$, Figure 4a in Appendix G illustrates ε-FSD, dashed areas represent the violation set.

Almost SSD (ε-SSD) We define ε-SSD, for $\varepsilon \in (0, \frac{1}{2})$, by relaxing Equation (5) as follows: $X \underset{\varepsilon-\text{SSD}}{\succcurlyeq} Y$ if and only if

$$\varepsilon_{IQ}(F_X, F_Y) = \frac{\int_0^1 (F_Y^{(-2)}(t) - F_X^{(-2)}(t))_+^2 dt}{d_{IQ}^2(F_X, F_Y)} \leq \varepsilon, \tag{8}$$

where d_{IQ} is the L_2 distance between the Integrated Quantiles ($F^{(-2)}$). This ratio corresponds to a measure of the "area" of violation of the SSD dominance of X on Y. Figure 4b in Appendix G illustrates the second order, dashed areas represent the violation set of SSD of X on Y. Appendix D gives a more detailed account on almost stochastic dominance.

2.3. Relative Stochastic Dominance

In the remainder of the paper, we refer to the FSD violation ratio as $\varepsilon_{\mathsf{W}_2}(F_X, F_Y) \equiv \varepsilon^{(1)}(F_X, F_Y)$ and to the SSD violation ratio as $\varepsilon_{IQ}(F_X, F_Y) \equiv \varepsilon^{(2)}(F_X, F_Y)$. One of the shortcomings of almost stochastic dominance is the need to fix a threshold ε on the violation ratio. When comparing two random variables, setting a threshold is a viable option. Nevertheless, when one needs to rank multiple variables $X_1, \ldots, X_k$ (considering all pairwise comparisons), setting a single threshold that would lead to a consistent relative stochastic dominance among the k variables becomes challenging. To alleviate this issue, we draw inspiration from relative similarity and dependence tests (Bounliphone et al., 2016a;b) that circumvent the need for a threshold via relative pairwise testings.

For $\ell \in \{1, 2\}$ (i.e for FSD or SSD) we consider all pairs of violations ratios:

$$\varepsilon_{ij}^{(\ell)} = \varepsilon^{(\ell)}(F_{X_i}, F_{X_j}) \text{ for } i, j \in \{1 \ldots k\}, i \neq j,$$

noting that $\varepsilon_{ij}^{(\ell)} + \varepsilon_{ji}^{(\ell)} = 1$. Let $F = (F_{X_1}, \ldots F_{X_k})$. We define the one-versus-all violation ratio of the dominance of

X_i on all other variables $X_j, j \neq i$:

$$\varepsilon_i^{(\ell)}(F) = \frac{1}{k-1}\sum_{j\neq i}\varepsilon_{ij}^{(\ell)}.$$

We then define relative stochastic dominance for both orders, R-FSD an R-SSD respectively:

$$\begin{aligned} &X_{i_1} \underset{R-\text{FSD}}{\succcurlyeq} X_{i_2} \cdots \underset{R-\text{FSD}}{\succcurlyeq} X_{i_k} \\ &\iff \varepsilon_{i_1}^{(1)}(F) \leq \cdots \leq \varepsilon_{i_k}^{(1)}(F) \qquad \text{and,} \\ &X_{i_1} \underset{R-\text{SSD}}{\succcurlyeq} X_{i_2} \cdots \underset{R-\text{SSD}}{\succcurlyeq} X_{i_k} \\ &\iff \varepsilon_{i_1}^{(2)}(F) \leq \cdots \leq \varepsilon_{i_k}^{(2)}(F) \end{aligned}$$

In this definition of relative stochastic dominance, the most dominating model is the one with the lowest one-versus-all violation ratio and to test for relative dominance of X_i on X_j we can look at the following statistics:

$$\Delta\varepsilon_{ij}^{(\ell)}(F) = \varepsilon_i^{(\ell)}(F) - \varepsilon_j^{(\ell)}(F), \tag{9}$$

and we have the following threshold-free test for relative order:[4]

$$X_i \underset{R-\text{FSD}}{\succcurlyeq} X_j \iff \Delta\varepsilon_{ij}^{(1)}(F) \leq 0 \tag{10}$$

$$X_i \underset{R-\text{SSD}}{\succcurlyeq} X_j \iff \Delta\varepsilon_{ij}^{(2)}(F) \leq 0 \tag{11}$$

3. Testing For Almost and Relative Stochastic Dominance

Given empirical samples from F_X and F_Y we perform statistical testing of the almost and relative stochastic dominance of X on Y given empirical estimates of the statistics given in Sections 2.2 and 2.3. A key ingredient for quantifying the statistical significance of such tests is a central limit theorem that guarantees that the centered empirical statistics is asymptotically Gaussian at the limit of infinite sample size. Given n samples from F_X (m from F_Y respectively), we denote F_X^n and F_Y^m the corresponding empirical distributions. For ε_0- FSD, (Del Barrio et al., 2018) studied the following hypothesis testing $H_0 : X \underset{\varepsilon_0-\text{SSD}}{\not\succcurlyeq} Y$ versus the alternative $H_a : X \underset{\varepsilon_0-\text{SSD}}{\succcurlyeq} Y$. Using (2), this amounts to the following null hypothesis : $H_0 : \varepsilon_{\mathsf{W}_2}(F_X^n, F_Y^m) > \varepsilon_0$. (Del Barrio et al., 2018) showed the asymptotic normality of the empirical statistics: (Del Barrio et al., 2018; Ulmer et al., 2022) propose to reject H_0 with a confidence level $1-\alpha$ if:

$$\varepsilon_{\mathsf{W}_2}(F_X^n, F_Y^m) \leq \varepsilon_0 + \sqrt{\frac{m+n}{mn}}\sigma^2(F_X, F_Y)\Phi^{-1}(\alpha), \tag{12}$$

where Φ^{-1} is the quantile function of a standard normal.

For the tests we propose below, we assume the following structure on the underlying CDFs to derive the corresponding central limit theorems (CLTs).

Assumption 1 (Regularity). Let the CDF F be supported on the interval $[-M, M]$ for some constant M, and have pdf f such that $\frac{f'(p)}{f^3(p)}$ is bounded for almost every p for which $f(p) > 0$ (i.e. all p in the support of f).

ε-SSD Testing Similar to ε-FSD, using the definition in (5) we propose to test using the following null hypothesis for testing for ε_0-SSD:

$$H_0 : \varepsilon_{IQ}(F_X^n, F_Y^m) > \varepsilon_0$$

Supposing Assumption 1 holds for F_X, F_Y and assuming $\frac{n}{n+m} \to \lambda$ for some λ, we state a Central Limit Theorem for the second order statistics (Theorem 3.1, proved in Appendix J.1).

Theorem 3.1 (Central Limit Theorem for ε-SSD)**.** *Assume that F_X, F_Y are supported on intervals*[5] *in $[-M, M]$, and have pdfs f_x, f_y such that $\frac{f_x'(p)}{f_x^3(p)}$, $\frac{f_y'(p)}{f_y^3(p)}$ are bounded almost everywhere on the support of f_x and f_y respectively. Assume we have n samples from F_X and m samples from F_Y, with $n, m \to \infty$ such that $\frac{n}{n+m} \to \lambda$ for some λ. Then $\sqrt{\frac{mn}{m+n}}\left(\varepsilon_{IQ}(F_X^n, F_Y^m) - \varepsilon_{IQ}(F_X, F_Y)\right) \to \mathcal{N}(0, \sigma_\lambda^2(F_X, F_Y))$ where $\sigma_\lambda^2(F_X, F_Y) = \frac{1}{d_{IQ}^8(F_X, F_Y)}\left[(1-\lambda)\text{Var}(v_X(U)) + \lambda\text{Var}(v_Y(U))\right]$, for $U \sim \text{Unif}[0,1]$, $v_Y(t) = 2\left(\frac{1}{f_y(F_Y^{-1}(t))}\right)\left(\int_t^1 (F_X^{(-2)}(p) - F_Y^{(-2)}(p))_+ dp\right)$, and $v_X(t) = 2\left(\frac{1}{f_x(F_X^{-1}(t))}\right)\left(\int_t^1 (F_X^{(-2)}(p) - F_Y^{(-2)}(p))_- dp\right)$.*

Similarly to (12), Theorem 3.1 suggests to reject H_0 with a confidence $1-\alpha$ if :

$$\varepsilon_{IQ}(F_X^n, F_Y^m) \leq \varepsilon_0 + \sqrt{\frac{m+n}{mn}}\sigma_\lambda^2(F_X, F_Y)\Phi^{-1}(\alpha), \tag{13}$$

where (for the same reasons as the FSD case) σ_λ^2 is given by the central limit theorem.

Relative Stochastic Dominance Testing We turn now to relative stochastic dominance that we introduced in (10)

[4] For comparing $k = 2$ random variables, these r-FSD and r-SSD tests reduce to 0.5-FSD and 0.5-SSD absolute tests, respectively.

[5] The interval for F_X and for F_Y need not coincide.

and (11) for first and second orders. Given n samples from k random variables $(X_1 \dots X_k)$, let $F = (F_1, \dots, F_k)$ be the marginals of X_i and $F_n = (F_{1n}, \dots, F_{kn})$ denote the empirical marginals. To test for R-FSD (resp R-SSD) of X_{i_1} on X_{i_2} we propose to test the following null hypothesis:

$$H_0 : \Delta\varepsilon^{(\ell)}_{ij}(F_n) > 0, \ell = 1 \text{ or } 2$$

Assuming that each F_i satisfies Assumption 1, we state in Appendix H a central limit theorem for the relative second order statistics (Theorem H.3 proved in in Appendix J.2). A similar result holds for the relative first order statistics that we omit for brevity. Theorem H.3 suggests to reject H_0 with a confidence $1 - \alpha$ if:

$$\Delta\varepsilon^{(2)}_{i_1,i_2}(F_n) \leq \sqrt{\frac{1}{n}\sigma^2_{relative}(F_X, F_Y)}\Phi^{-1}(\alpha) \tag{14}$$

where $\sigma^2_{relative}(F_X, F_Y)$ is given by the central limit theorem (similar test exists for R-FSD).

Bootstrapping Heuristic While the CLT above provides an asymptotic value for the variance, in practice (as in the ASO framework of (Ulmer et al., 2022)) we estimate the variance with a bootstrapping heuristic (Efron & Tibshirani, 1993). This estimate is nonasymptotic and hence should often be more accurate than the asymptotic value. Proving the consistency of the bootstrap for functions of quantiles is generally nontrivial (Shao & Tu, 2012), but recall that the stochastic ordering can be defined in terms of either quantiles or CDFs. In Appendix K we provide a bootstrap consistency proof for the absolute statistics based on the CDF, leaving the quantile based proof for future work.

Multi-Testing Algorithm Algorithm 1 given in Appendix C summarizes the multi-testing setup for both relative and almost (absolute) FSD and SSD. The main idea behind Algorithm 1 is to turn multi-testing to pairwise testings i.e testing for stochastic dominance between all pairs of models using relative (or absolute) FSD or SSD. In order to ensure that this multi-testing has a confidence level $1 - \alpha$, we correct the individual test's confidence level by dividing α by the number of all pairs (Bonferroni, 1936). Then in order to combine the pairwise rankings to a single rank, we use a simple Borda count (de Borda, 1781) rank aggregation algorithm.

4. Distributional Risk Aware Benchmarking of Foundation Models

Setup In this section we consider the multi-metric evaluation setup of a foundation model $A : \mathcal{X} \to \mathcal{O}$, using N metrics $m_i : \mathcal{O} \to \mathbb{R}, i = 1 \dots N$, where m_i are real valued functions evaluated on a test set D.Without loss of generality, assume that each of the metrics are standardized such that higher values of m_i correspond to more desirable model performance. We model observed values for each metric m_i as a continuous random variable M_i with unknown CDF F_{M_i}. For a model $A : \mathcal{X} \to \mathcal{O}$ and a data sample $X \sim D$, we describe the evaluation of model A with m_i with the following random variable M_i: $M_i|A, X := m_i(A(X)),\ X \sim D, i = 1 \dots N$, where the randomness arises from the data sampling procedure $X \sim D$, and (if applicable) the stochasticity of the model A, for example if the model uses sampling.

Metrics Portfolio Aggregation and Selection using Stochastic Dominance Let $\lambda = (\lambda_1, \dots, \lambda_N)$ be a probability vector that represents the importance of the m_i metrics to the model's end user. Inspired by the portfolio optimization literature, we model the user return from a model as a *portfolio of metrics m_i evaluated on a test set D*. Following (Ulan et al., 2021; Belgodere et al., 2023), we define this portfolio as an Independent copula, which forms a weighted geometric mean of the CDFs:

$$R_A(X) = \exp\left(\sum_{i=1}^{N} \lambda_i \log F_{M_i}\left(m_i(A(X))\right)\right) \tag{15}$$

Note that (15) normalizes the metrics using the CDF of the metric M_i, eliminating the issue of differing dynamic ranges. This CDF should be formed by pooling together the evaluations on all samples and from all models being compared, to ensure that the various R_A are comparable. The CDF normalization is monotonic and hence it preserves the order of each metrics and allow us to aggregate in the probability space the metrics using a simple weighted geometric mean. Computing $R_A(X)$ for all test samples X, we can therefore characterize the distribution of the metric portfolio of the model A. To compare two models it is enough to compare their corresponding portfolios, specifically, Model A is preferred to Model B using ε- or R-SSD:

$$R_A(X) \underset{\varepsilon- \text{ or } R-\text{SSD}}{\succcurlyeq} R_B(X). \tag{16}$$

Similar tests can be performed for FSD.

Note that the portfolio aggregation in (15) does not take into account the dependencies and correlations between the metrics. To alleviate this, we explore using also the empirical copula (Ruschendorf, 1976) as a means of aggregation of the metrics as follows

$$R^c_A(X) = \hat{C}\left(F_{M_1}\left(m_1(A(X))\right), \dots F_{M_N}\left(m_N(A(X))\right)\right), \tag{17}$$

where $\hat{C}$ is the empirical copula . Given N samples X_ℓ, $\ell = 1 \dots n$, the empirical copula is given by $\hat{C}(u_1, \dots u_n) = \frac{1}{n}\sum_{j=1}^{n}\Pi_{i=1}^{N}\mathbb{1}_{F_{M_i}(m_i(A(X_j)))<u_i}$. The empirical copula

can be understood as an average mean win rate (with an "and" operation on all metrics), that is computed on the CDF transformed scores of each evaluated sample. The main advantage of the independent copula (IC) in (15) versus the empirical copula (EC) in (17) is its computational efficiency ($O(nN)$ for IC versus $O(n^2N)$ for EC).

Multiple Models Comparison Given k models $A_\ell, \ell = 1 \dots k$ and their evaluations $m_i(A_\ell(X))$, $X \sim D$, $i = 1 \dots N$, we pool all model evaluations for a metric to estimate the CDF of each metric F_{M_i} and construct a portfolio for each model $R_{A_\ell}(X)$. We use our Relative Stochastic Dominance testing introduced in Section 3 and in Algorithm 1 to rank the models by their metrics portfolio in relative SSD or FSD with a confidence level $1 - \alpha$.

Per Metric Stochastic Dominance and Rank Aggregation We also explore another approach for multi-testing, by considering the stochastic dominance of the models on per-metric basis. This amounts to computing N relative stochastic orders for each $\mathcal{M}_i = (m_i(A_1(X)), \dots, m_i(A_\ell(X)))$, $i = 1 \dots N$. This amounts to producing via Algorithm 1 a relative ranking π_i of the models based on $\mathcal{M}_i$. A single rank π is then obtained via rank aggregation with uniform weighting on the per-metric rankings $\pi_i, i = 1 \dots N$. We use for rank aggregation the R package of (Pihur et al., 2009). For more details on rank aggregation, the reader is referred to Appendix F.3.

5. Experiments

5.1. Validation of Statistical Significance

We examine the statistical properties of our tests as a function of sample size. We purposely design synthetic score distributions to represent challenging problems comprising large overlap between the distributions and considerable violation ratio, but where one would still like to have an ordering among the variables. For this we consider the two Gaussian variables $X \sim \mathcal{N}(0, 1)$ and and $Y \sim \mathcal{N}(0.5, 2)$. Figure 5 in Appendix L.1 shows that our tests have desirable statistical properties. We perform synthetic experiment on fat tailed distribution such as log normal (Fig. 6 App. L.1).

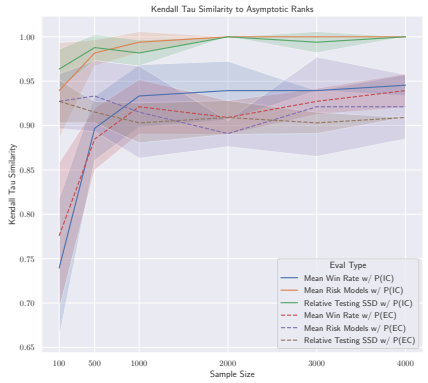
5.2. LLM evaluation with Stochastic Dominance

We showcase LLM evaluation with stochastic dominance to benchmark two risks: drifting from instructions and outputting toxic content. The following datasets correspond to each risk we benchmark.

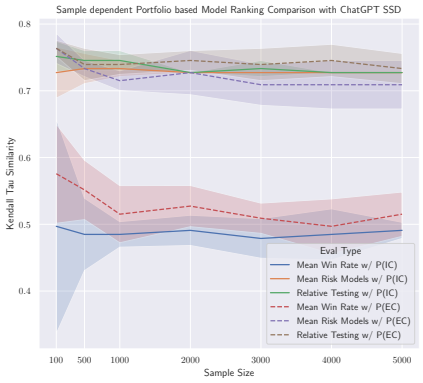
Mix-Instruct Evaluation Data We use the data from (Jiang et al., 2023), that consists of an instruction, an input sentence and an expected output from the user, as well as the output of a set of different LLMs. The dataset consists of a training set of 100K samples and a test set of 5K samples. (Jiang et al., 2023) used automatic metrics such as BARTscore and BLEU score comparing the LLM generation to the expected output in order to evaluate if each LLM followed the instruction. (Jiang et al., 2023) used also chatGPT to evaluate the generations (See Appendix B for ChatGPT evaluation). The number of automatic metrics N is 8, the total number of evaluated models k is 12. Metrics are unified so that larger values are preferred.

Toxicity Evaluation We use the real toxicity prompts dataset of Gehman et al. (2020), and generate prompts completions from the Llama 2 7b , Llama 2 13b, Llama 2 70b , MosaicML MPT 30b and Tiiuae Falcon 40b models available in Opensource ($k = 5$ models). We select two sets of prompts: toxic prompts (toxicity > 0.8, that gives $\sim$10K prompts) and non-toxic prompts (toxicity < 0.2, from which we randomly sample 10K). We sample from each model, 10 completions per prompt using nucleus sampling (top-p sampling with $p = 0.9$ and a temperature of 1). This procedure yields a dataset of $\sim$200K sentence completions per model. We evaluate the toxicity of these generations using the Perspective API, on the following toxicity metrics ($N = 6$ metrics): Toxicity, Severe toxicity, Identity Attack, Insult, Profanity and Threat. Following Liang et al. (2022), we evaluate the toxicity of generated completions only and refer to this as ***Gen Only*** evaluation. In order to also give the context of the completion, we prepend the model generation with the prompt and evaluate the full sentence using Perspective API. We refer to this as ***Prompt+Gen***. The polarity of all toxicity metrics is unified so that high values refer to non toxic content (we use $-\log$ probabilities of Perspective API outputs).

Evaluation Protocol and Baselines We evaluate each of the use cases (instruction following and toxicity) using the following absolute stochastic dominance tests: (1) ε-FSD (corresponds to the ASO evaluation of (Ulmer et al., 2022)) for $\varepsilon = 0.08, 0, 25, 0.4$. (2) our proposed ε-SSD using the same values for ε, (3) our relative stochastic dominance R-FSD and R-SSD tests, (4) the Mean – Risk models described in Table 3, and (5) the ranking produced by the Mean Win Rate (**MWR**) used by LLM leaderboards such as HELM (Liang et al., 2022). As noted in Section 4, we either perform these tests on a *metrics portfolio* – we refer to this as **test @P(IC)** when using the independent copula given in Equation (15) and **test @P(EC)** when using the empirical copula given in Equation (17) ; or on a per metric basis leading to N rankings of the models that we reduce to a single ranking via Rank Aggregation (RA) (Pihur et al., 2009) – we refer to this as **RA(test @ M)**. In this naming convention, **test** takes values in {MWR, ε-FSD, ε-SSD, R-FSD, R-SSD, Mean – Risk Model $(\mu_X - r_X)$} where r_X is a chosen risk from Table 3. We perform all our statistical tests with a significance level $\alpha = 0.05$, and use 1000



(a) Asymptotic Rank Stability



(b) Rank Similarity to R-SSD @chatGPT Rank.

Figure 2: (a) On the Mix-instruct dataset, we compute the ranking resulting from each ranking method using varying sample sizes from 100 to 5K. We repeat each experiment 5 times. We report for each method, the Kendall-Tau similarity between resulting ranks at each sample to the corresponding asymptotic rank at 5K samples. We see that Relative SSD on independent copula portfolio P(IC) is more stable in sample size than rank aggregation of all Mean Risk Models and more stable than MWR on the portfolio. The empirical dependent copula portfolio P(EC) does not have favorable asymptotics w.r.t to P(IC) since it suffers from the curse of dimension. (b) We use the same setup as in (a) but instead of Kendall-Tau similarity to the asymptotic rank of each method, we plot the similarity to R-SSD @ChatGPT rank at 5K samples. We see that MWR is inconsistent with chatGPT rank while both R-SSD @P(IC) and (EC) and RA(MRM @P(IC)) have a Kendall-Tau similarity between 0.7 and 0.75. Interestingly, the dependent copula (EC) captures better chatGPT rank than independent copula (IC), hinting at the favorable role of the metric dependencies.

bootstrap iterations.

Efficient Implementation We compare the computational complexity of our implementation for computing all stochastic orders to that of the `Deep-Significance` package (deepsig, 2022) which implements ε-FSD in the ASO framework (Ulmer et al., 2022), on the task of comparing models on the Mix-Instruct dataset (sample size 5K, $k = 12$ models). Using the `Deep-Significance` implementation of MULTI-ASO in (Ulmer et al., 2022) for $\varepsilon = 0.25$ with just 3 bootstrap iterations[6], the test completes in 15min50s (averaged over 7 runs). Our code for relative and absolute testing performs all tests at once and relies on caching vectorization and multi-threading of the operations. Our code completes all tests in an average of just 17.7 s with 1000 bootstraps. Experiments were run on a CPU machine with 128 AMD cores, of which 2 were used.

Mix-Instruct Results and Analysis In Figure 2 we depict the asymptotics of the ranks resulting from our tests as function of the sample size. In Figure 2 (a), we see that R-SSD with the portfolio aggregation with Independent Copula P(IC) has favorable asymptotics compared to R-SSD with dependent Empirical Copula P(EC). Indeed the empirical copula estimation suffers from the curse of dimension. On the other hand, we see in Figure 2 (b) that R-SSD with P(EC) captures better than P(IC) the ranks resulting from R-SSD with ChatGPT score. In other words, the dependent copula agrees more with the human evaluation proxy that is chatGPT. Note that the EC is expensive to compute and requires on average 1.5 hours on 5K samples, whereas IC requires only 0.87 seconds.

When compared with Mean Win Rate (MWR) used in LLM leaderboards such as HELM (Liang et al., 2022), we see that it does not have good asymptotics nor agree with ChatGPT rankings, regardless of the aggregation technique used. This is due to the fact that MWR only counts wins and does not take into account how fat is the left tail of the distribution of the metric being benchmarked, possibly leading to overevaluation of risky models.

Remarkably, the R-SSD ordering agrees with the rank aggregation of all (consistent) mean – risk models, confirming the theoretical link between second order dominance and risk averse decision making. The dependent copula EC with R-SSD leads to a better agreement with chatGPT R-SSD ranking than MRM models. Finally Tables 4 and Table 5 in Appendix L give additional results on R-FSD and the rank aggregation of all metrics, and how it compares to $\varepsilon-$ FSD and SSD.

Toxicity Results and Analysis Table 2 shows the results of our tests on the combined set of toxic and non toxic prompts. Ablation studies on individual sets are given in Table 6 in Appendix L.4. We make a few observations: First, overall

[6]Limited to 3 for computational reasons.

Scenario	Llama 2 7b	Llama 2 13b	Llama 2 70b	MosaicML MPT 30b	Tiiuae Falcon 40b
All Combined (Toxic + Non-Toxic Prompts)					
RA(R-FSD @M) (Gen Only)	**2**	**3**	5	**1**	4
R-FSD @P(IC) (Gen Only)	**2**	**3**	5	**1**	4
RA(R-SSD @M) (Gen Only)	**2**	**3**	5	**1**	4
R-SSD @P(IC) (Gen Only)	**2**	**3**	5	**1**	4
RA(R-FSD @M) (Prompt + Gen)	**3**	4	5	**1**	**2**
RA(R-FSD @M) (Prompt + Gen)	**3**	4	5	**1**	**2**
R-SSD @P(IC) (Prompt + Gen)	**3**	4	5	**1**	**2**
R-SSD @P(IC) (Prompt + Gen)	**3**	4	5	**1**	**2**

Table 2: Toxicity Ranking using an Independent Copula portfolio aggregation of Perspective API metrics.

the portfolio with independent copula approach agrees well with the rank aggregation of per-metric rankings. The portfolio is more computationally efficient as it needs to run the stochastic dominance test only on the portfolio, rather than running N tests and aggregating them via rank aggregation. An ablation study on empirical copula in Appendix L shows that it leads to a similar ranking as the Independent Copula. Secondly, on this dataset the R-FSD and R-SSD agree, with a few exceptions. Interestingly, when comparing models on model generation only, on toxic prompts MosaicML MPT stands out, while on non toxic prompts Llama2 7B stands out and on the combined set Mosaic ML MPT stands out. On the combined set, we see for the llama family that increased model size increases the toxicity of generations. This is in line with findings in the recent TrustLLM benchmark (Sun et al., 2024).

6. Conclusion

In this paper we introduced a distributional framework for risk aware benchmarking and comparison of foundation models based on multi-metric evaluations. Our framework has potential beyond the current applications presented here, being applicable wherever statistical significance while ranking assets for decision making is needed. We believe our tools for training models to be risk averse can be of significant use to practitioners and serve as a stepping stone towards solving the AI alignment problem.

Impact Statement

This paper presents a risk aware framework for benchmarking LLMs. In benchmarking LLM the stochastic nature of their generation and in presence of multiple metrics to be evaluated, our work offers a solution that gives raise to 1) a sound aggregation of the metrics via the copula method 2) a risk aware evaluation that takes into account tail events of misalignment and not only the average behaviors thanks to the use of stochastic orders 3) quantifies the uncertainty of the evaluation via statistical significance testings. The potential societal consequences of our work falls under AI governance as it allows a rigorous certification of compliance of LLMs with multitude of safeguards and dimensions.

References

Pedro C Alvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A contamination model for approximate stochastic order: extended version. *arXiv preprint arXiv:1412.1920*, 2014.

Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navratil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, Jerret Ross, Yair Schiff, Radhika Vedpathak, and Richard A. Young. Auditing and generating synthetic data with controllable trust trade-offs, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 2023.

C.E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber, 1936. URL `https://books.google.com/books?id=3CY-HQAACAAJ`.

Wacha Bounliphone, Eugene Belilovsky, Matthew Blaschko, Ioannis Antonoglou, and Arthur Gretton. A test of relative similarity for model selection in generative models. *Proceedings ICLR 2016*, 2016a.

Wacha Bounliphone, Eugene Belilovsky, Arthur Tenenhaus, Ioannis Antonoglou, Arthur Gretton, and Matthew B Blashcko. Fast non-parametric tests of relative dependency and similarity. *arXiv preprint arXiv:1611.05740*, 2016b.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

Jean-Charles de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

deepsig. Deepsignificance. `https://github.com/Kaleidophon/deep-significance`, 2022.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. An optimal transportation approach for assessing almost stochastic order. *The Mathematics of the Uncertain: A Tribute to Pedro Gil*, pp. 33–44, 2018.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392, 2018.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance-how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2773–2785, 2019.

B. Efron and R. Tibshirani. An Introduction to the Bootstrap, 1993.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings*, 2020. URL `https://api.semanticscholar.org/CorpusID:221878771`.

Alexander A Gushchin and Dmitriy A Borzykh. Integrated quantile functions: properties and applications. *Modern Stochastics: Theory and Applications*, 4(4):285–314, 2017.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *arXiv preprint arXiv:2309.07462*, 2023.

Yue Huang, Qihui Zhang, Lichao Sun, et al. Trustgpt: A benchmark for trustworthy and responsible large language models. *arXiv preprint arXiv:2306.11507*, 2023.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023.

Moshe Leshno and Haim Levy. Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance. *Management Science*, 48(8):1074–1085, 2002.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2022.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

Wlodzimierz Ogryczak and Andrzej Ruszczynski. Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization*, 13(1):60–78, 2002.

Vasyl Pihur, Susmita Datta, and Somnath Datta. Rankaggreg, an r package for weighted rank aggregation. *BMC Bioinformatics*, 10:62 – 62, 2009. URL `https://api.semanticscholar.org/CorpusID:206970248`.

Ludger Ruschendorf. Asymptotic Distributions of Multivariate Rank Order Statistics. *The Annals of Statistics*, 4 (5):912 – 923, 1976.

Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

Edwin D Simpson. Statistical significance testing for natural language processing, 2021.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Trustllm: Trustworthiness in large language models, 2024.

L.Y. Tzeng, Rachel Huang, and P.T. Shih. Revisiting almost second-degree stochastic dominance. *Management Science*, 59:1250–1254, 05 2013. doi: 10.2307/23443939.

Maria Ulan, Welf Löwe, Morgan Ericsson, and Anna Wingkvist. Copula-based software metrics aggregation. *Software Quality Journal*, 29(4):863–899, 2021. URL https://doi.org/10.1007/s11219-021-09568-9.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. deepsignificance-easy and meaningful statistical significance testing in the age of neural networks. *arXiv preprint arXiv:2204.06815*, 2022.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models, 2023.

Edward Beeching , Clémentine Fourrier , Nathan Habib , Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. Summit: Iterative text summarization via chatgpt. *arXiv preprint arXiv:2305.14835*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

Supplementary Material

A. Ablation Studies

Metrics Aggregation Versus Portfolio For portoflio, computing ranking using FSD and SSD including the portfolio computation on $5K$ samples for 5 bootstrap samples , we have mean execution time of 32.01 ± 4.51 s. For FSd and SSD ranking computation for all metrics, followed by rank using pearson distance the execution time is of 254.99 ± 16.76 s. On the other hand, we observe on the mix-instruct dataset a consistency of ranks between these two approaches (FSD or SDD on portfolio & FSD or SSD on all metrics followed by rank aggregation) as quantified by the kendall-tau similarity between the ranks:

1. Kendall Tau(R-SSD@P(IC), RA(R-SSD@M)) = 0.848
2. Kendall Tau(R-FSD@P(IC), RA(R-FSD@M)) = 0.878
3. Kendall Tau(R-SSD@P(EC), RA(R-SSD@M)) = 0.848
4. Kendall Tau(R-FSD@P(EC), RA(R-SSD@M)) = 0.848

We see that these two approaches lead to similar ranks while portfolio approach leads to 7x speedups when using IC portfolios.

B. Transforming Discrete Relative ChatGPT Scores to Absolute Real Valued Scores

We follow (Jiang et al., 2023) in mapping discrete chatGPT scores to real valued ones. Note that chatGPT scores for comparing models A and B are discrete and are one of these 4 options: A is better, B is better, Both are equally good, Both are equally bad.

Given m models we construct for each prompt sample $\ell = 1 \dots N$ a $m \times m$ comparison matrix with chatGPT:

$$X_{\ell,ij} = +1, X_{\ell,ji} = -1 \text{ if model } i \text{ is better}$$

$$X_{\ell,ij} = -1, X_{\ell,ji} = +1 \text{ if model } j \text{ is better}$$

$$X_{\ell,ij} = X_{\ell,ji} = +0.5 \text{ if model } i \text{ and } j \text{ equally good}$$

$$X_{\ell,ij} = X_{\ell,ji} = -0.5 \text{ if model } i \text{ and } j \text{ equally bad}$$

Then each model will define the following scalar score at each sample ℓ:

$$s_{\ell,i} := \sum_{j=1}^{m} (X_{\ell,ij} - X_{\ell,ji}).$$

hence we have a distribution of chatGPT score for each model :

$$p_i = \frac{1}{N} \sum_{i=1}^{N} \delta_{s_{\ell,i}}, i = 1 \dots m.$$

Note that the scores $s_{\ell,i}$ take on even integer values between $-2m$ and $2m$ inclusive, we treat the support as continuous and consider the following kernel density estimator with Gaussian kernel of width σ:

$$\hat{p}_i^{(\sigma)}(t) = \frac{1}{N} \sum_{i=1}^{N} \varphi \left(\frac{t - s_{\ell,i}}{\sigma} \right), t \in \mathbb{R},\ i = 1 \dots m,$$

where φ is the standard normal density. In Figure 3 below we plot chatGPT scores kernel density estimates for two models, openassistant and flan-t5:

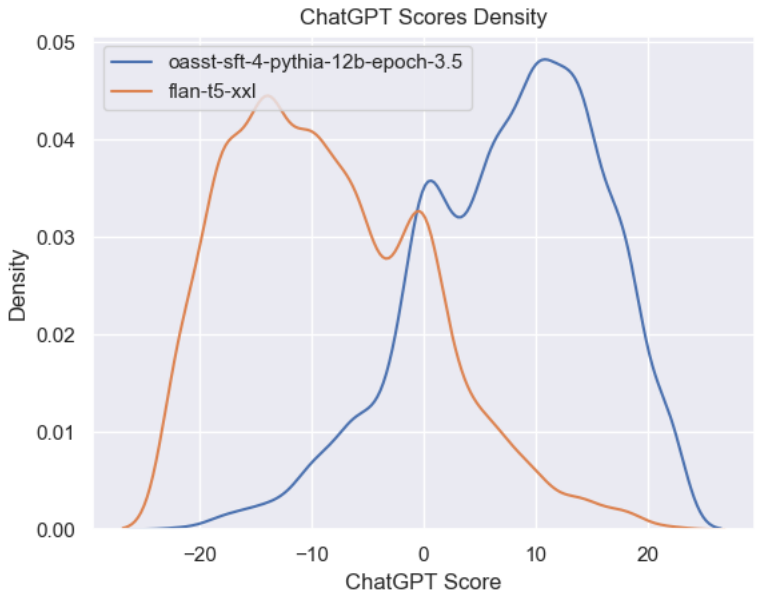


Figure 3: ChatGPT density scores for two models, open-assistant has clearly higher scores than the Flan-t5 models.

C. Multi-Testing Algorithm for Relative and Almost Stochastic Dominance

Our multi-testing algorithm for relative and almost stochastic dominance is detailed in Algorithm 1.

In a nutshell our multi-testing consists of the following steps:

1. For evaluation of each model compute summary statistics , i.e quantiles and integrated quantiles.
2. For all pairs of models, compute statistics of absolute and relative tests by computing violation ratios.
3. Compute the variance of these statistics via bootstrapping.
4. Perform the hypothesis testing for all pairs models with a corrected confidence level taking into account the number of all pairs
5. Aggregate pairwise rankings to a rank using the Borda algorithm, that ranks the model by their number of wins in the stochastic dominance tests performed above.

D. Absolute or Almost Stochastic Dominance

Almost FSD (ε-FSD) Following (Leshno & Levy, 2002), (Del Barrio et al., 2018) relaxed FSD via the violation ratio of FSD:

Definition D.1 (FSD Violation Ratio (Del Barrio et al., 2018))**.** For $F_X \neq F_Y$ define the violation ratio:

$$\varepsilon_{\mathsf{W}_2}(F_X, F_Y) = \frac{\int_{A_0^{(1)}} (F_X^{(-1)}(t) - F_Y^{(-1)}(t))^2 dt}{\int_0^1 (F_X^{(-1)}(t) - F_Y^{(-1)}(t))^2 dt} = \frac{\int_0^1 (F_Y^{(-1)}(t) - F_X^{(-1)}(t))_+^2 dt}{\mathsf{W}_2^2(F_X, F_Y)},$$

where $A_0^{(1)} = \left\{ t \in (0,1) : F_Y^{(-1)}(t) > F_X^{(-1)}(t) \right\}$ is the violation set the relation $X \underset{\text{FSD}}{\succcurlyeq} Y$, and W_2 is the Wasserstein-2 distance.

Algorithm 1 Stochastic Order Multi-testing (relative and **absolute**)

Input: $F_1, ..., F_k$, k models we want to rank corresponding to empirical measure $p_1 = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^1}, \dots p_k = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i^k}$, **Threshold:** τ.
Input: Desired stochastic order $\in \{1, 2\}$, B number of bootstraps, $m = K^2$ number of comparisons, significance level α.
Cache the bootstraps samples and their statistics
for $j = 1$ **to** k **do**
  $p_j^0 \leftarrow p_j$
  Get Quantiles and Integrated Quantiles
  $\mathrm{Q}_{0,j} \leftarrow \textsc{GetQuantiles}(p_j)$
  $\mathrm{IQ}_{0,j} \leftarrow \textsc{GetIntegratedQuantiles}(p_j)$
  for $b = 1$ **to** B **do**
    Get Quantiles and Integrated Quantiles
    $p_j^b \leftarrow \textsc{ResampleWithReplacement}(p_j, n)$ {using quantiles and uniform}
    $\mathrm{Q}_{b,j} \leftarrow \textsc{GetQuantiles}(p_j^b)$
    $\mathrm{IQ}_{b,j} \leftarrow \textsc{GetIntegratedQuantiles}(p_j^b)$
  end for
end for
Compute all violation ratios
$\varepsilon_{b,i,j} \leftarrow \textsc{ComputeViolationRatios}(F_i^b, F_j^b, \text{order})$ for $b = 0 \dots B, \quad i, j = 1 \dots k, i \neq j$ {ratio of Q or IQ of $j > i$ by total area}
$\varepsilon_{b,i,i} = 0, \forall\, b, i$
Compute the sum statistics
for $b = 0$ **to** B **do**
  for $i = 1$ **to** k **do**
    $\varepsilon_b^i \leftarrow \frac{1}{k-1}\sum_j \varepsilon_{b,i,j}$
  end for
end for
Compute the relative statistics
$\Delta\varepsilon_b^{i,j} = \varepsilon_b^i - \varepsilon_b^j, \forall b, i, j$
Compute the Bootstrap Variance
for $i = 1$ **to** k **do**
  for $j = 1$ **to** k **do**
    $\sigma_{ij} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\Delta\varepsilon_b^{i,j} - \textsc{Mean}(\Delta\varepsilon_b^{i,j}, b))^2}$
    $\sigma_{ij}^{\text{abs}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\varepsilon_{b,i,j} - \textsc{Mean}(\varepsilon_{b,i,j}, b))^2}$
  end for
end for
Compute the test
$\mathrm{Win}_{ij} = \mathrm{Win}_{ij}^{\text{abs}} = 0$
for $i = 1$ **to** k **do**
  for $j = 1$ **to** k **do**
    if $i \neq j$ and $\Delta\varepsilon_0^{i,j} - \frac{1}{\sqrt{n}}\sigma_{ij}\Phi^{-1}(\alpha/k^2) \leq 0$ **then**
      $\mathrm{Win}_{ij} = 1$ {with confidence level $1 - \alpha/k^2$}
    end if
    if $i \neq j$ and $\varepsilon_{0.i,j} - \frac{1}{\sqrt{n}}\sigma_{ij}^{\text{abs}}\Phi^{-1}(\alpha/k^2) \leq \tau$ **then**
      $\mathrm{Win}_{ij}^{\text{abs}} = 1$ {with confidence level $1 - \alpha/k^2$}
    end if
  end for
end for
rank $= \textsc{Borda}(\mathrm{Win})$ {with confidence level $1 - \alpha$}
$\mathrm{rank}_{\text{abs}} = \textsc{Borda}(\mathrm{Win}^{\text{abs}})$ {with confidence level $1 - \alpha$}
Return rank, $\mathrm{rank}_{\text{abs}}$

Algorithm 2 COMPUTEVIOLATIONRATIOS(F_a,F_b,order)

if order =1 **then**
  Return $\varepsilon_{\mathsf{W}_2}(F_a, F_b)$ in Definition D.1
else if order=2 **then**
  Return $\varepsilon_{IQ}(F_a, F_b)$ in Definition D.2
end if

Note that $0 \leq \varepsilon_{\mathsf{W}_2}(F_X, F_Y) \leq 1$, with value 0 if $X \underset{\text{FSD}}{\succ} Y$ and 1 if $Y \underset{\text{FSD}}{\succ} X$. For $\varepsilon \in (0, \frac{1}{2}]$, the relaxed FSD can be therefore defined as follows

$$X \underset{\varepsilon-\text{FSD}}{\succcurlyeq} Y \iff \varepsilon_{\mathsf{W}_2}(F_X, F_Y) \leq \varepsilon. \tag{18}$$

Figure 4a in Appendix G illustrates ε-FSD, dashed areas represent the violation set.

Almost SSD (ε-SSD) Note that the original definition of ε-FSD of X on Y in (Leshno & Levy, 2002) is an L_1 definition and uses the CDF rather than quantiles: $\int_{-\infty}^{\infty}(F_X(x) - F_Y(x))_+ dx \leq \varepsilon \int_{\infty}^{\infty} |F_X(x) - F_Y(x)| dx$. (Tzeng et al., 2013) gave a similar L_1 definition for ε-SSD using the second performance function $F^{(2)}(.)$. According to (Tzeng et al., 2013), X dominates Y in the ε-SSD if $\int_{-\infty}^{\infty}(F_X^{(2)}(x) - F_Y^{(2)}(x))_+ dt \leq \varepsilon \int_{-\infty}^{+\infty} |F_X^{(2)}(x) - F_Y^{(2)}(x)| dx$. Following (Del Barrio et al., 2018), we redefine ε-SSD using second quantiles and with a L_2 definition, this eases the analysis and practically the integration is on $(0, 1]$ rather than $(-\infty, \infty)$.

We define the SSD violation ratio as follows:

Definition D.2 (SSD Violation Ratio)**.** For $F_X \neq F_Y$ define the violation ratio:

$$\varepsilon_{IQ}(F_X, F_Y) = \frac{\int_{A_0^{(2)}}(F_X^{(-2)}(t) - F_Y^{(-2)}(t))^2 dt}{\int_0^1 (F_X^{(-2)}(t) - F_Y^{(-2)}(t))^2 dt} = \frac{\int_0^1 (F_Y^{(-2)}(t) - F_X^{(-2)}(t))_+^2 dt}{d_{IQ}^2(F_X, F_Y)},$$

where $A_0^{(2)} = \left\{t \in (0,1) : F_Y^{(-2)}(t) > F_X^{(-2)}(t)\right\}$ is the violation set the relation $X \underset{\text{SSD}}{\succcurlyeq} Y$, and d_{IQ} is the L_2 distance between the Integrated Quantiles ($F^{(-2)}$).

We are now ready to define ε-SSD, for $\varepsilon \in (0, \frac{1}{2})$:

$$X \underset{\varepsilon-\text{SSD}}{\succcurlyeq} Y \iff \varepsilon_{IQ}(F_X, F_Y) \leq \varepsilon \tag{19}$$

Figure 4b in Appendix G illustrates the second order, dashed areas represent the violation set of SSD of X on Y. Integrated quantiles fully characterize one dimensional distributions as can be seen from the Theorem I.1 stated and proved in Appendix I:

E. Related Works on Stochastic Dominance

Stochastic Dominance In (Dror et al., 2018; 2019; Ulmer et al., 2022; Simpson, 2021) a distributional assessment of the models based on stochastic dominance was proposed to overcome the limitations of the ubiquitous Mean-Variance Risk model used in machine learning.

(Ulmer et al., 2022) used first order almost stochastic dominance and advocated for selecting a model A over B based on a metric m_i if: $M_i|A, X \underset{\varepsilon-\text{FSD}}{\succcurlyeq} M_i|B, X$. We expand this to the Relative-FSD. In natural language (and other) applications, it is often crucial to mitigate the risk of outputs with low metrics, especially when those metrics quantify important socio-technical guardrails such as model's toxicity, safety, or robustness. Unfortunately, the first stochastic ordering does

not capture an assessment of the left tail behavior of $M_i|A, X$ and $M_i|B, X$ and hence does not provide a risk-aware benchmarking (Ogryczak & Ruszczynski, 2002). To alleviate this issue, we instead consider the *second* order stochastic ordering and use our second order *almost* or *relative* stochastic dominance tests introduced in Section 3 for selecting a model A if:$M_i|A, X \underset{\varepsilon \text{ or } R-\text{SSD}}{\succcurlyeq} M_i|B, X$.

F. Supplement Discussions

F.1. Mean Risk Models

Name	Risk Measure	$\alpha-$ consistency with SSD
Standard deviation	$\sigma_X = \sqrt{\mathbb{E}(X-\mu_X)^2}$	not consistent
Absolute semi deviation	$\delta_X = \mathbb{E}(\mu_X - X)_+$	$1-$ consistent
Negative Tail Value at Risk	$-\text{TVAR}_X(p) = -\frac{F^{(-2)}(p)}{p}$	$1-$ consistent for all $p \in (0,1]$
Mean absolute deviation from a quantile	$h_X(p) = \mu_x - \frac{F_X^{(-2)}(p)}{p}$	$1-$ consistent for all $p \in (0,1]$
Gini Tail	$\Gamma_X = 2\int_0^1(\mu_X p - F_X^{(-2)}(p))dp$	$1-$ consistent

Table 3: Risk models and their $\alpha-$consistency with SSD.

Note that several risks in Table 3 use the second quantile function as part of a benchmarking of the left tails of the outcomes.

F.2. $\delta-$ Consistency of Gini-Risk Models with ε-SSD

$\delta-$ Consistency of Gini-Risk Models with ε-SSD We relax the definition of $\alpha-$ consistency of mean-risk models with SSD to (α, δ) consistency with ε-SSD as follows:

Definition F.1 ((α, δ) consistency of MRM with ε-SSD)**.** A mean-risk model (μ_X, r_X) is (α, δ) consistent with ε-SSD, if there exists $\alpha, \delta > 0$ such that $X \underset{\varepsilon\text{-SSD}}{\succcurlyeq} Y \implies \mu_X - \alpha r_x + \delta \geq \mu_Y - \alpha r_Y$

It is easy to see that the Mean-Gini tail MRM of X and Y is consistent with their ε-SSD:

Proposition F.2. *The Mean-Gini Tail MRM is* $(1, 2\varepsilon^{\frac{1}{2}} d_{IQ}(F_X, F_Y))$ *consistent with* ε*-SSD.*

Proof of Proposition F.2.

$$
\begin{aligned}
\mu_X - \Gamma_X &= \mu_X - 2\int_0^1(\mu_X p - F_X^{(-2)}(p))dp = 2\int_0^1(F_X^{(-2)}(p) - F_Y^{(-2)}(p) + F_Y^{(-2)}(p))dp \\
&= 2\int_0^1 F_Y^{(-2)}(p) + 2\int_{A_0^{(2)}}(F_X^{(-2)}(p) - F_Y^{(-2)}(p))dp + 2\underbrace{\int_{[0,1]/A_0^{(2)}}(F_X^{(-2)}(p) - F_Y^{(-2)}(p))dp}_{\geq 0} \\
&\geq 2\int_0^1 F_Y^{(-2)}(p) - 2\int_{A_0^{(2)}}|F_X^{(-2)}(p) - F_Y^{(-2)}(p))|dp \\
&= \mu_Y - \Gamma_Y - 2\int_0^1(F_Y^{(-2)}(p) - F_X^{(-2)}(p))_+dp \\
&\geq \mu_Y - \Gamma_Y - 2\left(\int_0^1 dp\right)^{\frac{1}{2}}\left(\int_0^1(F_Y^{(-2)}(p) - F_X^{(-2)}(p))_+^2 dp\right)^{\frac{1}{2}} \text{ (Cauchy-Schwartz)} \\
&\geq \mu_Y - \Gamma_Y - 2\varepsilon^{\frac{1}{2}} d_{IQ}(F_X, F_Y) \text{(By assumption } X \underset{\varepsilon-\text{SSD}}{\succcurlyeq} Y\text{)}
\end{aligned}
$$

□

F.3. Rank Aggregation

Given N ranks $\pi_i, i = 1 \dots N$ represented as permutations in S_k, the rank aggregation in (Pihur et al., 2009) solves the following problem :

$$\min_{\pi \in S_k} \sum_{i=1}^{N} \alpha_i d(\pi, \pi_i),$$

where $\alpha_i \geq 0, \sum_{i=1}^{N} \alpha_i = 1$ represent importance of each ranking and d is a distance between permutations. (Pihur et al., 2009) have multiple choices of distance such as Pearson or Kendall's-Tau. We fixed through out our experiments the distance to Pearson.

F.4. Mean Win Rate and CDF normalizers in portfolio

To unpack the notations in (15), consider a distribution $\mathcal{A}$ on models space. For a sample $X \sim D_i$ and a model $A \sim \mathcal{A}$, the metric $m_i()$ normalization through its CDF can be written as follows:

$$F_{M_i}(m_i(A(X)) = \mathbb{E}_{B \sim \mathcal{A}} \mathbb{E}_{Y \sim D_i} \mathbb{1}_{m_i(B(Y) \leq m_i(A(X))}. \tag{20}$$

Hence for a model A on each evaluated sample the CDF normalizer computes a soft ranking of the evaluation of the model A with a metric m_i on the sample X with respect to all models and all samples.

Remark F.3 (Mean Win Rate). Note that in LLM leaderborads such as HELM and Hugging face, the performance of a model A evaluated with a metric m_i, is summarized via a Mean Win Rate (MWR) aggregated on models level looking on expected metrics

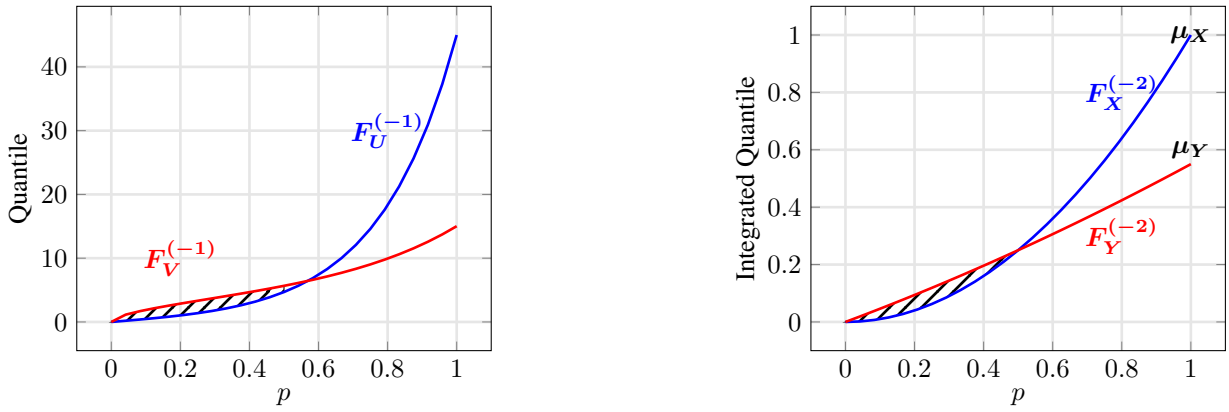
$$\mathrm{MWR}_{A,M_i} = \mathbb{E}_{B \sim \mathcal{A}} \mathbb{1}_{\mathbb{E}_{X \sim D_i}[m_i(B(X))] \leq \mathbb{E}_{X \sim D_i}[m_i(A(X))]}, \tag{21}$$

or aggregated on sample level marginalizing on models with a $\max$:

$$\overline{\mathrm{MWR}}_{A,M_i} = \mathbb{E}_{X \sim D_i} \mathbb{1}_{\max_{B \neq A} m_i(B(X)) \leq m_i(A(X))}, \tag{22}$$

Contrasting (20) , (21) and (22) we see that instead of looking at the MWR summary statistics that does not allow to consider all order statistics and relative ordering as well the risks of tails events, we consider a full distributional benchmarking in the metrics portfolio approach.

G. Figures



(a) ε- FSD (first order): $U \underset{\varepsilon-\mathrm{FSD}}{\succcurlyeq} V$

(b) ε-SSD (second order): $X \underset{\varepsilon-\mathrm{SSD}}{\succcurlyeq} Y$

Figure 4: **(a) An Example of Almost First Order Stochastic Dominance:** Plots of quantile functions of U and V. Dashed areas is the violation set of first order stochastic dominance of U on V. **(b) An Example of Almost Second Order Stochastic Dominance:** Plots of integrated quantile functions; dashed area is the violation set for the second order stochastic dominance of X on Y.

H. Central Limit Theorems

H.1. CLT for ε-SSD

Theorem H.1 (Central Limit Theorem for ε-SSD)**.** *Assume that F_X, F_Y are supported on intervals*[a] *in $[-M, M]$, and have pdfs f_x, f_y such that $\frac{f_x'(p)}{f_x^3(p)}$, $\frac{f_y'(p)}{f_y^3(p)}$ are bounded almost everywhere on the support of f_x and f_y respectively. Assume we have n samples from F_X and m samples from F_Y, with $n, m \to \infty$ such that $\frac{n}{n+m} \to \lambda$ for some λ. Then*

$$\sqrt{\frac{mn}{m+n}}\left(\varepsilon_{IQ}(F_X^n, F_Y^m) - \varepsilon_{IQ}(F_X, F_Y)\right) \to \mathcal{N}(0, \sigma_\lambda^2(F_X, F_Y))$$

where

$$\sigma_\lambda^2(F_X, F_Y) = \frac{1}{d_{IQ}^8(F_X, F_Y)}\left[(1-\lambda)\mathrm{Var}(v_X(U)) + \lambda \mathrm{Var}(v_Y(U))\right],$$

for $U \sim \mathrm{Unif}[0,1]$, $v_Y(t) = 2\left(\frac{1}{f_y(F_Y^{-1}(t))}\right)\left(\int_t^1 (F_X^{(-2)}(p) - F_Y^{(-2)}(p))_+ dp\right)$, and $v_X(t) = 2\left(\frac{1}{f_x(F_X^{-1}(t))}\right)\left(\int_t^1 (F_X^{(-2)}(p) - F_Y^{(-2)}(p))_- dp\right)$.

[a] The interval for F_X and for F_Y need not coincide.

Remark H.2 (Non-independent samples). Theorem H.1 assumes that the n-sample from F_X is independent of the m-sample for F_Y. Consider instead the setting where there are n samples from F_X and F_Y that are dependent (e.g. X, Y are evaluations of different models applied to the same data). We can describe general dependence structure as the following. Suppose (X, Y) has marginals $X \sim F_X$, $Y \sim F_Y$, with some unknown dependence structure (optionally described by the copula $C_{XY}(u_x, u_y) = \Pr(F_X(X) \le u_x, F_Y(Y) \le u_y)$). Let

$$(U_x, U_y) = (F_X(X), F_Y(Y)) \sim C_{XY}.$$

Note that U_x and U_y have marginals equal to $\mathrm{Unif}([0,1])$, but U_x and U_y may be dependent. Hence the variances in each term of the decomposition (24) in the appendix cannot be added. Instead, one should modify the result of Theorem H.1 to use

$$\bar{\sigma}_\lambda^2(F_X, F_Y) = \frac{1}{d_{IQ}^8(F_X, F_Y)}\mathrm{Var}\left[v_X(U_x) + v_Y(U_y)\right].$$

H.2. CLT for Relative Statistics

We focus here on presenting the Central Limit Theorem for SSD. The relative FSD has a similar form and we omit its statement here.

Theorem H.3 (Central limit Theorem for Relative SSD). *Assume $F_{1n}, \ldots, F_{kn}$ are available and independent, and each F_i satisfies the conditions of Theorem H.1. Then*

$$\sqrt{n}\left(\Delta\varepsilon^{(2)}_{i_1,i_2}(F_n) - \Delta\varepsilon^{(2)}_{i_1,i_2}(F)\right) \to_w \mathcal{N}\left(0, \frac{1}{(k-1)^2}\sum_{i=1}^{k}\sigma_i^2(i_1,i_2)\right).$$

where

$$\sigma_i^2(i_1,i_2) = \begin{cases} \mathrm{Var}\left(\frac{2v^{(1)-}_{i_1i_2}(U_i)}{d^4_{IQ}(F_{i_1},F_{i_2})} + \sum_{j\neq i_1,i_2}\frac{v^{(1)-}_{i_1j}(U_i)}{d^4_{IQ}(F_{i_1},F_j)}\right) & i = i_1 \\ \mathrm{Var}\left(\frac{2v^{(2)+}_{i_1i_2}(U_i)}{d^4_{IQ}(F_{i_1},F_{i_2})} - \sum_{j\neq i_1,i_2}\frac{v^{(1)-}_{i_2j}(U_i)}{d^4_{IQ}(F_{i_2},F_j)}\right) & i = i_2 \\ \mathrm{Var}\left(\frac{v^{(2)+}_{i_1j}(U_i)}{d^4_{IQ}(F_{i_1},F_j)} - \frac{v^{(2)+}_{i_2j}(U_i)}{d^4_{IQ}(F_{i_2},F_j)}\right) & i \neq i_1, i_2 \end{cases}$$

for $U_i \sim \mathrm{Unif}([0,1])$ all independent, and $v^{(1),-}_{ij}(t) = 2\left(\frac{dF_i^{-1}(t)}{dt}\right)\left(\int_t^1 (F_i^{(-2)}(p) - F_j^{(-2)}(p))_- dp\right)$, $v^{(2),+}_{ij}(t) = 2\left(\frac{dF_j^{-1}(t)}{dt}\right)\left(\int_t^1 (F_i^{(-2)}(p) - F_j^{(-2)}(p))_+ dp\right)$.

Remark H.4 (Dependent samples). If the F_{in} are dependent, a similar expression to that shown in Remark H.2 for the absolute testing case also holds here. The statement is omitted.

I. Proof of Theorem I.1

Theorem I.1 (d_{IQ} is a metric). *d_{IQ} is a metric on the space of univariate distributions with continuous CDF, moreover, it metrizes the weak topology.*

First, we show that $d_{IQ}(F,G) = 0$ if and only if $F = G$. The forward direction is obvious. For the reverse direction, if $d_{IQ}(F,G) = 0$, then $F^{(-2)}(t) = G^{(-2)}(t)$ a.e. By the continuity of integrated quantiles, this implies $F^{(-2)} = G^{(-2)}$ everywhere. Then, since $F^{(-1)}(t)$ is simply the derivative of $F^{(-2)}(t)$ with respect to t[7], $F^{(-1)} = G^{(-1)}$ everywhere by differentiating both sides of $F^{(-2)}(t) = G^{(-2)}(t)$. Hence $F = G$ since distributions are uniquely determined by their quantile functions.

The triangle inequality follows from the triangle inequality of the L_2 norm, since $\sqrt{\int_0^1 (F^{(-2)}(t) - G^{(-2)}(t))^2 dt} = \|F^{(-2)}(t) - G^{(-2)}(t)\|_{L_2([0,1])}$. Hence d_{IQ} is a metric. By Theorem 10 in (Gushchin & Borzykh, 2017), we know that random variable $X_{(i)} \to_w X$ (with cdf $F_{(i)}$) if and only if $F^{(-2)}_{(i)}$ converges uniformly to $F^{(-2)}$. Hence d_{IQ} must metrize weak convergence.

J. Proofs of Central Limit Theorems

J.1. Absolute Testing: Proof of Theorem H.1

Note that for U_i and V_i an n-sample and an m-sample respectively from $\mathrm{Unif}([0,1])$, we can get X_i, Y_i as $X_i = F^{-1}(U_i)$, $Y_i = G^{-1}(V_i)$. Let $H_{n,1}$ and $H_{m,2}$ be the empirical d.f.s of the U_i and V_i respectively. We have

$$F_n^{-1}(t) = F^{-1}(H_{n,1}^{-1}(t)),$$

hence

$$F_n^{(-2)}(t) = \int_0^t F_n^{-1}(p)dp = \int_0^t F^{-1}(H_{n,1}^{-1}(p))dp.$$

[7] This follows because F^{-2} is the integral of the finite-valued quantile function $F^{-1}(t)$.

We are interested in

$$\varepsilon_{IQ}(F_n, G_m) = \frac{\int_{A_0}(F_n^{(-2)}(t) - G_m^{(-2)}(t))^2 dt}{d_{IQ}^2(F_n, G_m)},$$

where

$$A_0 = \left\{t \in (0,1) : G_m^{(-2)}(t) > F_n^{(-2)}(t)\right\},$$

is the violation set.

It is shown in (Gushchin & Borzykh, 2017) (Theorem 10 therein) that integrated quantiles converge uniformly, i.e. $F_n^{(-2)}(t) \to F^{(-2)}(t)$ pointwise. As an immediate consequence, we have

$$\varepsilon_{IQ}(F_n, G_m) \to_{a.s.} \varepsilon_{IQ}(F, G).$$

We apply the following decomposition and bound the two terms separately:

$$\varepsilon_{IQ}(F_n, G_m) - \varepsilon_{IQ}(F, G) = (\varepsilon_{IQ}(F_n, G_m) - \varepsilon_{IQ}(F, G_m)) + (\varepsilon_{IQ}(F, G_m) - \varepsilon_{IQ}(F, G)). \tag{23}$$

We derive asymptotic normality of these terms for G_m, the proof for F_n is identical by symmetry.

We introduce the statistics

$$S_m = \int_0^1 (F^{(-2)}(t) - G_m^{(-2)}(t))^2 dt$$

$$S_m^+ = \int_0^1 (F^{(-2)}(t) - G_m^{(-2)}(t))_+^2 dt$$

$$S_m^- = \int_0^1 (F^{(-2)}(t) - G_m^{(-2)}(t))_-^2 dt$$

The nonrandom S, S^+, S^- are defined similarly with G instead of G_m.

Next, set

$$T_m = \sqrt{m}(S_m - S)$$

$$T_m^+ = \sqrt{m}(S_m^+ - S^+)$$

$$T_m^- = \sqrt{m}(S_m^- - S^-).$$

Theorem J.1. *Assume that G is supported on an interval that is a subset of $[-M, M]$, and has pdf g such that $\frac{g'(p)}{g^3(p)}$ is bounded almost everywhere on the support of g. Then*

$$T_m = \alpha_{m,2}(v) + o_P(1)$$

$$T_m^+ = \alpha_{m,2}(v^+) + o_P(1)$$

$$T_m^- = \alpha_{m,2}(v^-) + o_P(1)$$

where we define $\alpha_{m,2}(t) = \sqrt{m}(t - H_{m,1}^{-1}(t))$ and $\alpha_{m,2}(v) = \int_0^1 v(t)\alpha_{m,2}(t)dt$, and

$$v(t) = 2\left(\frac{1}{g(G^{-1}(t))}\right)\left(\int_t^1 F^{(-2)}(p) - G^{(-2)}(p)dp\right).$$

$$v^+(t) = 2\left(\frac{1}{g(G^{-1}(t))}\right)\left(\int_t^1 (F^{(-2)}(p) - G^{(-2)}(p))_+ dp\right),$$

$$v^-(t) = 2\left(\frac{1}{g(G^{-1}(t))}\right)\left(\int_t^1 (F^{(-2)}(p) - G^{(-2)}(p))_- dp\right).$$

Proof. We begin with T_m. Note that[8]

$$\begin{aligned}
T_m &= \sqrt{m}(S_m - S) \\
&= \sqrt{m}\int_0^1 (F^{(-2)}(t) - G_m^{(-2)}(t))^2 - (F^{(-2)}(t) - G^{(-2)}(t))^2 dt \\
&= \sqrt{m}\int_0^1 \left[2F^{(-2)}(t) - G_m^{(-2)}(t) - G^{(-2)}(t)\right](G^{(-2)}(t) - G_m^{(-2)}(t))dt \\
&\to 2\sqrt{m}\int_0^1 \left[F^{(-2)}(t) - G^{(-2)}(t)\right](G^{(-2)}(t) - G_m^{(-2)}(t))dt \\
&= 2\sqrt{m}\int_0^1 \left[F^{(-2)}(t) - G^{(-2)}(t)\right]\left[\int_0^t G^{(-1)}(p) - G^{(-1)}(H_{m,1}^{-1}(p)))dp\right]dt
\end{aligned}$$

Let us do integration by parts:

$$\begin{aligned}
&2\sqrt{m}\int_0^1 \left[F^{(-2)}(t) - G^{(-2)}(t)\right]\left[\int_0^t G^{(-1)}(p) - G^{(-1)}(H_{m,1}^{-1}(p)))dp\right]dt = \\
&= 2\sqrt{m}\left[\left(\int_0^1 F^{(-2)}(t) - G^{(-2)}(t)dt\right)\left[\int_0^1 G^{(-1)}(t) - G^{(-1)}(H_{m,1}^{-1}(t)))dt\right]\right. \\
&\left. - \int_0^1 \left(\int_0^t F^{(-2)}(p) - G^{(-2)}(p)dp\right)\left[G^{(-1)}(t) - G^{(-1)}(H_{m,1}^{-1}(t)))\right]dt\right] \\
&= 2\sqrt{m}\int_0^1 \left(\int_t^1 F^{(-2)}(p) - G^{(-2)}(p)dp\right)\left[G^{(-1)}(t) - G^{(-1)}(H_{m,1}^{-1}(t)))\right]dt \\
&= 2\sqrt{m}\int_0^1 \left(\frac{dG^{-1}(t)}{dt}\right)\left(\int_t^1 F^{(-2)}(p) - G^{(-2)}(p)dp\right)(t - H_{m,1}^{-1}(t))dt \\
&+ O\left(\sqrt{m}\int_0^1\int_t^1 F^{(-2)}(p) - G^{(-2)}(p)dp)(t - H_{m,1}^{-1}(t))^2 dt\right) \\
&= 2\sqrt{m}\int_0^1 \left(\frac{dG^{-1}(t)}{dt}\right)\left(\int_t^1 F^{(-2)}(p) - G^{(-2)}(p)dp\right)(t - H_{m,1}^{-1}(t))dt + o_P(1).
\end{aligned}$$

In the penultimate step we have used a first-order Taylor series on $G^{-1}(t)$ via the assumption that $\frac{d^2G^{-1}(t)}{dt^2} = -\frac{g'(G^{-1}(t))}{g^3(G^{-1}(t))}$ is bounded almost everywhere, and in the final step we have noted that

$$\begin{aligned}
\sqrt{m}\int_0^1 \left(\int_t^1 F^{(-2)}(p) - G^{(-2)}(p)dp\right)(t - H_{m,1}^{-1}(t))^2 dt &\leq 2\sqrt{m}\int_0^1 (t - H_{m,1}^{-1}(t))^2 dt \\
&= o_P(1),
\end{aligned}$$

since the support of F and G lie in $[-M, M]$ and $\int_0^1 (t - H_{m,1}^{-1}(t))^2 dt = O_p(1/m)$.

We then have

$$T_m = \alpha_{m,2}(v) + o_P(1),$$

where $\alpha_{m,2}(t) = \sqrt{m}(t - H_{m,1}^{-1}(t))$, and $\alpha_{m,2}(v) = \int_0^1 v(t)\alpha_{m,2}(t)dt$ where

$$v(t) = 2\left(\frac{dG^{-1}(t)}{dt}\right)\left(\int_t^1 F^{(-2)}(p) - G^{(-2)}(p)dp\right).$$

Similarly,

$$T_m^+ = \alpha_{m,2}(v^+) + o_P(1), \quad T_m^- = \alpha_{m,2}(v^-) + o_P(1)$$

[8] Convergence here is uniform convergence of the integrated quantiles.

where

$$v^+(t) = 2\left(\frac{dG^{-1}(t)}{dt}\right)\left(\int_t^1 (F^{(-2)}(p) - G^{(-2)}(p))_+ dp\right),$$

$$v^-(t) = 2\left(\frac{dG^{-1}(t)}{dt}\right)\left(\int_t^1 (F^{(-2)}(p) - G^{(-2)}(p))_- dp\right).$$

□

Corollary J.2. *Assume that G is supported on an interval in $[-M, M]$, and has pdf g such that $\frac{g'(p)}{g^3(p)}$ is bounded almost everywhere on the support of g. Then as $m \to \infty$*

$$\sqrt{m}(\epsilon_{IQ}(F, G_m) - \epsilon_{IQ}(F, G)) \to_w \mathcal{N}(0, \sigma^2)$$

and if additionally $n \to \infty$

$$\sqrt{m}(\epsilon_{IQ}(F_n, G_m) - \epsilon_{IQ}(F_n, G)) \to_w \mathcal{N}(0, \sigma^2),$$

if for $U \sim \text{Unif}([0, 1])$

$$\sigma^2 = \frac{\text{Var}(v^+(U))}{d_{IQ}^8(F, G)}$$

is finite.

Proof. Note that by Theorem J.1

$$\sqrt{m}(\epsilon_{IQ}(F, G_m) - \epsilon_{IQ}(F, G)) = \sqrt{m}\left(\frac{S_m^-}{S_m} - \frac{S^-}{S}\right) = \frac{\sqrt{m}}{SS_m}(T_m^- - T_m) \to -\frac{\alpha_{m,2}(v^+)}{S^2}$$

since $S_m \to S$ a.s. Recalling the definition of $\alpha_{m,2}$ yields asymptotic normality with zero mean as in (Del Barrio et al., 2018), and variance as calculated in the corollary statement.

The case of $\sqrt{m}(\epsilon_{IQ}(F_n, G_m) - \epsilon_{IQ}(F_n, G))$ follows similarly since integrated quantiles weakly converge as $F_n \to F$. □

Continuing with the main proof, recalling (23) and using Corollary J.2 along with the asymptotic independence of the two terms and the fact that $\frac{n}{n+m} \to \lambda$, we have

$$\begin{aligned}\sqrt{\frac{mn}{m+n}}\,&(\varepsilon_{IQ}(F_n, G_m) - \varepsilon_{IQ}(F, G)) \\ &= \sqrt{(1-\lambda)n}(\varepsilon_{IQ}(F_n, G_m) - \varepsilon_{IQ}(F, G_m)) + \sqrt{\lambda n}(\varepsilon_{IQ}(F, G_m) - \varepsilon_{IQ}(F, G)) \\ &\to \mathcal{N}(0, \sigma_\lambda^2(F, G))\end{aligned} \tag{24}$$

where

$$\sigma_\lambda^2(F, G) = \frac{1}{d_{IQ}^8(F, G)}\left[(1-\lambda)\text{Var}(v_F(U)) + \lambda \text{Var}(v_G(U))\right].$$

Here, we have defined

$$v_G(t) = 2\left(\frac{1}{g(G^{-1}(t))}\right)\left(\int_t^1 (F^{(-2)}(p) - G^{(-2)}(p))_+ dp\right),$$

and

$$v_F(t) = 2\left(\frac{1}{f(F^{-1}(t))}\right)\left(\int_t^1 (F^{(-2)}(p) - G^{(-2)}(p))_- dp\right).$$

J.2. Relative Testing: Proof of Theorem H.3

Note that

$$\begin{aligned}\Delta\varepsilon_{IQ}^{i_1,i_2}(F) &= \epsilon_{IQ}^{i_1}(F) - \epsilon_{IQ}^{i_2}(F)\\ &= \frac{1}{k-1}\left[\sum_{j\neq i_1}\epsilon_{IQ}^{i_1 j} - \sum_{j\neq i_2}\epsilon_{IQ}^{i_2 j}\right]\\ &= \frac{1}{k-1}\left[2\epsilon_{IQ}^{i_1 i_2} - 1 + \sum_{j\neq i_1,i_2}(\epsilon_{IQ}^{i_1 j} - \epsilon_{IQ}^{i_2 j})\right].\end{aligned}$$

For compactness, let us introduce the differencing notation $\phi(\cdot)|_F^{F_n} = \phi(F_n) - \phi(F)$. We seek a CLT on

$$\begin{aligned}\sqrt{n}(\widehat{\Delta\varepsilon_{IQ}^{i_1,i_2}}(F_n) - \Delta\varepsilon_{IQ}^{i_1,i_2}(F)) &= \frac{\sqrt{n}}{k-1}\left(2\epsilon_{IQ}(\cdot,F_{i_2,n}) + \sum_{j\neq i_1,i_2}\epsilon_{IQ}(\cdot,F_{j,n})\right)\Bigg|_{F_{i_1}}^{F_{i_1,n}}\\ &\quad + \frac{\sqrt{n}}{k-1}\left(2\epsilon_{IQ}(F_{i_1},\cdot) - \sum_{j\neq i_1,i_2}\epsilon_{IQ}(\cdot,F_{j,n})\right)\Bigg|_{F_{i_2}}^{F_{i_2,n}}\\ &\quad + \frac{\sqrt{n}}{k-1}\sum_{j\neq i_1,i_2}(\epsilon_{IQ}(F_{i_1},\cdot) - \epsilon_{IQ}(F_{i_2},\cdot))|_{F_j}^{F_{j,n}}\\ \to_w \underbrace{\frac{\sqrt{n}}{k-1}\left(2\epsilon_{IQ}(\cdot,F_{i_2}) + \sum_{j\neq i_1,i_2}\epsilon_{IQ}(\cdot,F_j)\right)\Bigg|_{F_{i_1}}^{F_{i_1,n}}}_{I} &+ \underbrace{\frac{\sqrt{n}}{k-1}\left(2\epsilon_{IQ}(F_{i_1},\cdot) - \sum_{j\neq i_1,i_2}\epsilon_{IQ}(\cdot,F_j)\right)\Bigg|_{F_{i_2}}^{F_{i_2,n}}}_{II}\\ &\quad + \underbrace{\frac{\sqrt{n}}{k-1}\sum_{j\neq i_1,i_2}(\epsilon_{IQ}(F_{i_1},\cdot) - \epsilon_{IQ}(F_{i_2},\cdot))|_{F_j}^{F_{j,n}}}_{III}\end{aligned}$$

where we have used the uniform convergence of integrated quantiles. Note that I, II, and each term in the sum in III are all independent.

Define

$$\begin{aligned}v_{ij}^{(1)}(t) &= 2\left(\frac{dF_i^{-1}(t)}{dt}\right)\left(\int_t^1 F_i^{(-2)}(p) - F_j^{(-2)}(p)dp\right),\\ v_{ij}^{(2)}(t) &= 2\left(\frac{dF_j^{-1}(t)}{dt}\right)\left(\int_t^1 F_i^{(-2)}(p) - F_j^{(-2)}(p)dp\right),\end{aligned}$$

and $v_{ij}^{(1)+}$, $v_{ij}^{(2)+}$ similarly. Then by the proof of Corollary J.2, each term in III converges to

$$\begin{aligned}\frac{\sqrt{n}}{k-1}(\epsilon_{IQ}(F_{i_1},\cdot) - \epsilon_{IQ}(F_{i_2},\cdot))|_{F_j}^{F_{j,n}} &\to -\frac{\alpha_{m,j}(v_{i_1 j}^{(2)+})}{(k-1)d_{IQ}^4(F_{i_1},F_j)} + \frac{\alpha_{m,j}(v_{i_2 j}^{(2)+})}{(k-1)d_{IQ}^4(F_{i_2},F_j)}\\ &= \frac{1}{k-1}\alpha_{m,j}\left(-\frac{v_{i_1 j}^{(2)+}}{d_{IQ}^4(F_{i_1},F_j)} + \frac{v_{i_2 j}^{(2)+}}{d_{IQ}^4(F_{i_2},F_j)}\right)\\ &\to_w \mathcal{N}\left(0, \frac{1}{(k-1)^2}\sigma_j^2(i_1,i_2)\right), \quad \forall j\neq i_1,i_2.\end{aligned}$$

where

$$\sigma_j^2(i_1,i_2) = \frac{1}{(k-1)^2}\mathrm{Var}\left(\frac{v_{i_1 j}^{(2)+}(U)}{d_{IQ}^4(F_{i_1},F_j)} - \frac{v_{i_2 j}^{(2)+}(U)}{d_{IQ}^4(F_{i_2},F_j)}\right), \quad \forall j\neq i_1,i_2,$$

and $U \sim \text{Unif}([0,1])$. Similarly for I and II,

$$I \to_w \mathcal{N}\left(0, \frac{1}{(k-1)^2}\sigma^2_{i_1}(i_1,i_2)\right)$$
$$II \to_w \mathcal{N}\left(0, \frac{1}{(k-1)^2}\sigma^2_{i_2}(i_1,i_2)\right)$$

where[9]

$$\sigma^2_{i_1}(i_1,i_2) = \text{Var}\left(\frac{2v^{(1)-}_{i_1 i_2}(U)}{d^4_{IQ}(F_{i_1},F_{i_2})} + \sum_{j\neq i_1,i_2}\frac{v^{(1)-}_{i_1 j}(U)}{d^4_{IQ}(F_{i_1},F_j)}\right),$$
$$\sigma^2_{i_2}(i_1,i_2) = \text{Var}\left(\frac{2v^{(2)+}_{i_1 i_2}(U)}{d^4_{IQ}(F_{i_1},F_{i_2})} - \sum_{j\neq i_1,i_2}\frac{v^{(1)-}_{i_2 j}(U)}{d^4_{IQ}(F_{i_2},F_j)}\right).$$

Putting everything together via independence,

$$\sqrt{n}\left(\widehat{\Delta\varepsilon^{i_1,i_2}_{IQ}}(F_n) - \Delta\varepsilon^{i_1,i_2}_{IQ}(F)\right) \to_w \mathcal{N}\left(0, \frac{1}{(k-1)^2}\sum_{i=1}^{k}\sigma^2_i(i_1,i_2)\right).$$

K. Consistency of Bootstrapping

In this section, we consider the relaxation measure using the CDFs[10]:

$$\tilde{\epsilon}_\ell(F_X,F_Y) = \frac{\int_{-\infty}^{\infty}(F_Y^{(\ell)}(t) - F_X^{(\ell)}(t))^2_+ dt}{\int_{-\infty}^{\infty}(F_Y^{(\ell)}(t) - F_X^{(\ell)}(t))^2 dt}.$$

Note that we can relax FSD as follows:

$$Y \underset{\varepsilon-\text{FSD}}{\succcurlyeq} X \iff \tilde{\epsilon}_1(F_X,F_Y) \leq \varepsilon. \tag{25}$$

Similarly we can relax SSD as follows:

$$Y \underset{\varepsilon-\text{SSD}}{\succcurlyeq} X \iff \tilde{\epsilon}_2(F_X,F_Y) \leq \varepsilon. \tag{26}$$

We will prove bootstrap consistency for $\ell = 1$ (approximate first order dominance), the proof for $\ell = 2$ (approximate second order dominance) is similar.

We seek to show that the bootstrapped variance $\mathsf{Var}(\tilde{\epsilon}_1(F_X^{n*},F_Y^{m*}))$ is an asymptotically consistent estimator of $\mathsf{Var}(\tilde{\epsilon}_1(F_X^n,F_Y^m)$, i.e. their ratio goes to 1:

$$\frac{\mathsf{Var}(\tilde{\epsilon}_1(F_X^{n*},F_Y^{m*}))}{\mathsf{Var}(\tilde{\epsilon}_1(F_X^n,F_Y^m))} \to_p 1.$$

Note we can write this as

$$\frac{\mathsf{Var}(\tilde{\epsilon}_1(F_X^{n*},F_Y^{m*})}{\mathsf{Var}(\tilde{\epsilon}_1(F_X^n,F_Y^m))} \to_p \frac{\mathsf{Var}(T(F_X^{n*},F_Y^{m*}))}{\mathsf{Var}(T(F_X^n,F_Y^m))},$$

where

$$T(F_X,F_Y) = \frac{\int_{-\infty}^{\infty}(F_Y(t) - F_X(t))^2_+ dt}{\int_{-\infty}^{\infty}(F_Y(t) - F_X(t))^2 dt}.$$

[9] This $U \sim \text{Unif}([0,1])$ is drawn simply for this variance calculation and is not dependent on anything outside of this equation.
[10] The result using quantiles as described in the main text is less straightforward and if left for future work.

Consider the metric created by the sup norm

$$\rho_\infty(F,G) = \|F-G\|_\infty = \sup_x |F(x)-G(x)|.$$

Note that T is continuously ρ_∞-Frechet differentiable in both arguments due to the differentiability of the function $(\cdot)_+^2$ and integration. Specifically,

$$\begin{aligned} D_{1,(F_X,F_Y)}(G_X) = &\frac{1}{(\int_{-\infty}^{\infty}(F_Y(t)-F_X(t))^2dt)^2}\cdot \\ &\Big[\left(\int_{-\infty}^{\infty}(F_Y(t)-F_X(t))^2dt\right)\left(\int_{-\infty}^{\infty}2(F_Y(t)-F_X(t))_+G_X dt\right) \\ &-\left(\int_{-\infty}^{\infty}(F_Y(t)-F_X(t))_+^2dt\right)\left(\int_{-\infty}^{\infty}2(F_Y(t)-F_X(t))G_X dt\right)\Big]. \end{aligned}$$

and similarly for $D_{2,(F_X,F_Y)}(G_Y)$. Since T is continuously differentiable, by the definition of continuous Frechet differentiability we can write (see Chapter 2 in (Shao & Tu, 2012)) the following:

$$\begin{aligned} &T(F_X^{n*},F_Y^{m*}) - T(F_X^n,F_Y^m) \\ &= D_{1,(F_X^n,F_Y^m)}(F_X^{n*}-F_X^n) + D_{2,(F_X^n,F_Y^m)}(F_Y^{m*}-F_Y^m) + (\rho_\infty(F_X^{n*},F_X^n)+\rho_\infty(F_Y^{m*},F_Y^m))\epsilon_{n,m}^* \end{aligned}$$

$$T(F_X^{n*},F_Y^m) - T(F_X^n,F_Y^m) = D_{1,(F_X^n,F_Y^m)}(F_X^{n*}-F_X^n) + (\rho_\infty(F_X^{n*},F_X^n))\epsilon_n^*$$

$$T(F_X^n,F_Y^{m*}) - T(F_X^n,F_Y^m) = D_{2,(F_X^n,F_Y^m)}(F_Y^{m*}-F_Y^m) + (\rho_\infty(F_Y^{m*},F_Y^m))\epsilon_m^*$$

and

$$\begin{aligned} &T(F_X^n,F_Y^m) - T(F_X,F_Y) \\ &= D_{1,(F_X,F_Y)}(F_X^n-F_X) + D_{2,(F_X,F_Y)}(F_Y^m-F_Y) + (\rho_\infty(F_X^n,F_X)+\rho_\infty(F_Y^m,F_Y))\epsilon_{n,m} \end{aligned}$$

$$T(F_X^n,F_Y) - T(F_X,F_Y) = D_{1,(F_X,F_Y)}(F_X^n-F_X) + (\rho_\infty(F_X^n,F_X))\epsilon_n$$

$$T(F_X,F_Y^m) - T(F_X,F_Y) = D_{2,(F_X,F_Y)}(F_Y^m-F_Y) + (\rho_\infty(F_Y^m,F_Y))\epsilon_m$$

where $\epsilon_{n,m}^*, \epsilon_n^*, \epsilon_m^*, \epsilon_{n,m}, \epsilon_n, \epsilon_m \to 0$ as $n,m \to \infty$.

Hence, combining terms,

$$T(F_X^{n*},F_Y^{m*}) - T(F_X^n,F_Y^m) = (T(F_X^{n*},F_Y^m) - T(F_X^n,F_Y^m)) + (T(F_X^n,F_Y^{m*}) - T(F_X^n,F_Y^m)) + o_p(n^{-1/2}+m^{-1/2}),$$

and

$$T(F_X^n,F_Y^m) - T(F_X,F_Y) = (T(F_X^n,F_Y) - T(F_X,F_Y)) + (T(F_X,F_Y^m) - T(F_X,F_Y)) + o_p(n^{-1/2}+m^{-1/2}).$$

Hence, assuming independence of the n-sample and m-sample and respective bootstrap resamplings,

$$\frac{\mathsf{Var}(T(F_X^{n*},F_Y^{m*}))}{\mathsf{Var}(T(F_X^n,F_Y^m))} \to_{a.s.} \frac{\mathsf{Var}(T(F_X^n,F_Y^{m*})) + \mathsf{Var}(T(F_X^{n*},F_Y^m))}{\mathsf{Var}(T(F_X,F_Y^m)) + \mathsf{Var}(T(F_X^n,F_Y))},$$

i.e. we add the variances.

We have now divided the task to the one-sided setting where the bootstrap is only done in one argument of T. Hence we can apply Theorem 3.10 of (Shao & Tu, 2012) which states that for ρ_∞-Frechet differentiable functions of a CDF, the bootstrap variance estimator is asymptotically consistent if the support is bounded (more general results can be stated but are omitted for simplicity). Applying separately to each of the two variances we have the following.

Proposition K.1. *Suppose F_X, F_Y, have support contained in $[-M,M]$ for some $M>0$, and F_X^n, F_Y^m arise from independent samples. Then*

$$\frac{\mathsf{Var}(\tilde{\epsilon}_1(F_X^{n*},F_Y^{m*}))}{\mathsf{Var}(\tilde{\epsilon}_1(F_X^n,F_Y^m))} \to_{a.s.} 1.$$

L. Additional Experimental Results

L.1. Statistical Significance on Synthetic Data

We examine the statistical properties of our tests as a function of sample size. We purposely design synthetic score distributions to represent challenging problems with large overlap between the distributions and a considerable violation ratio, but where one would still like to have an ordering among the variables. For this we consider the two Gaussian distributions with mean $\mu = 0$ and standard deviation $\sigma = 1$, and with mean $\mu = 0.5$ and standard deviation $\sigma = 2$, respectively. In the top panels of Figure 5 we show the PDF, CDF and integrated quantile function of these two Gaussians, illustrating the large violation ratio. The orange distribution can be calculated to be 0.2-FSD and 0.45-SSD over the blue distribution. Note that these ε values are not comparable, due to the differences in definitions. In Figure 5, we conduct experiments illustrating the power of our tests for the absolute tests of the hypotheses $H_{0,FSD} = 0.45$-FSD and $H_{0,SSD} = 0.45$-SSD. We also use our relative tests, which in this 2-variable case (as noted in the main text) are equivalent to testing $H_{0,FSD} = 0.5$-FSD and $H_{0,SSD} = 0.5$-SSD. The bottom left panel in Figure 5 show the True Positive Rate for the different types of tests that we developed: relative test with quantile function, relative test with Integrated Quantile Function, absolute test with quantile function, and absolute test with Integrated Quantile Function. As expected, all tests quickly converge towards True Positive Rate of 1.0 for growing sample sizes.

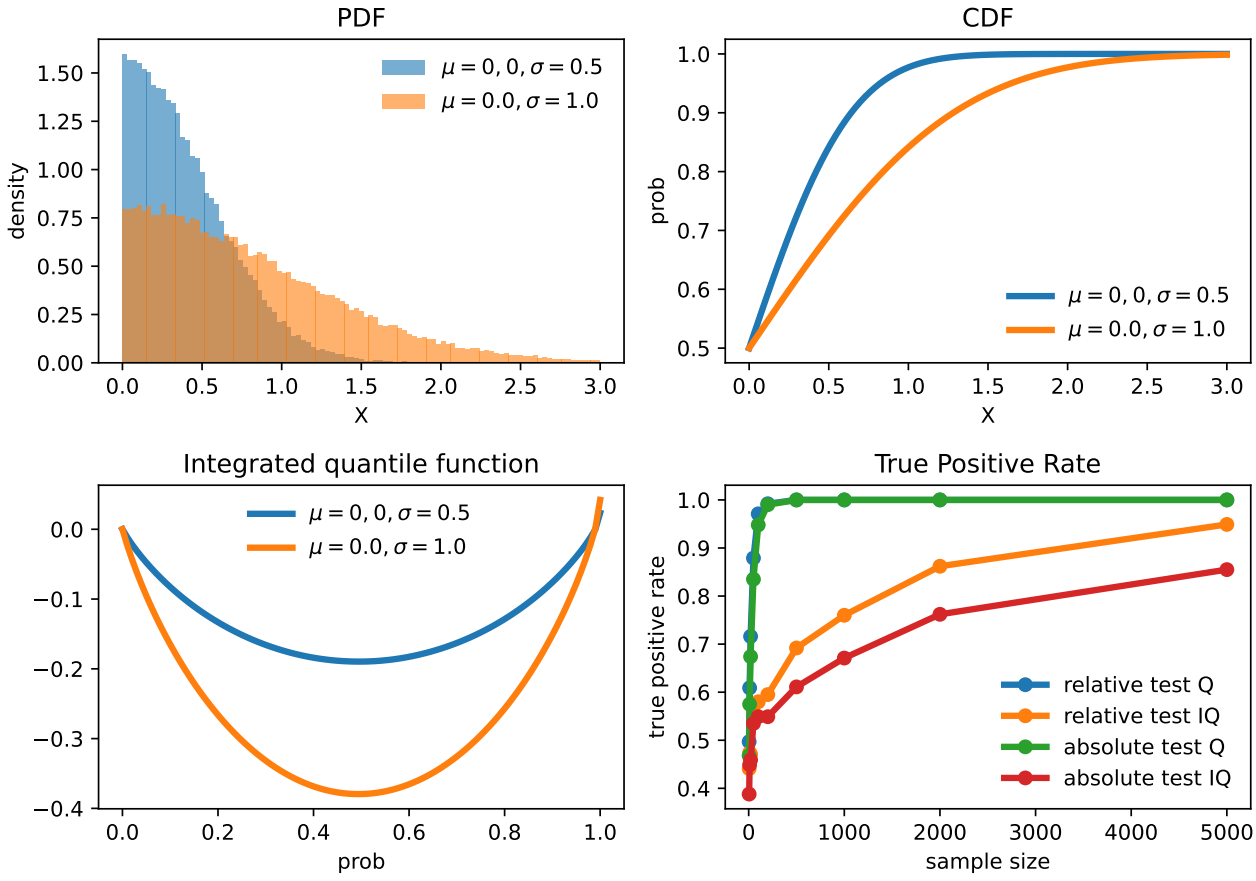


Figure 5: True Positive Rate vs sample size for **Gaussian distributions**. We compute the True Positive Rate of our stochastic dominance methods on the test distributions in the top panels for different sample sizes. Decisions are made using a confidence threshold of $\alpha = 0.05$ and $\tau = 0.45$ (for the absolute tests) and rates are computed over 1000 repetitions of the tests. Note that the FSD and SSD curves should not be compared due to differences in the underlying hypotheses.

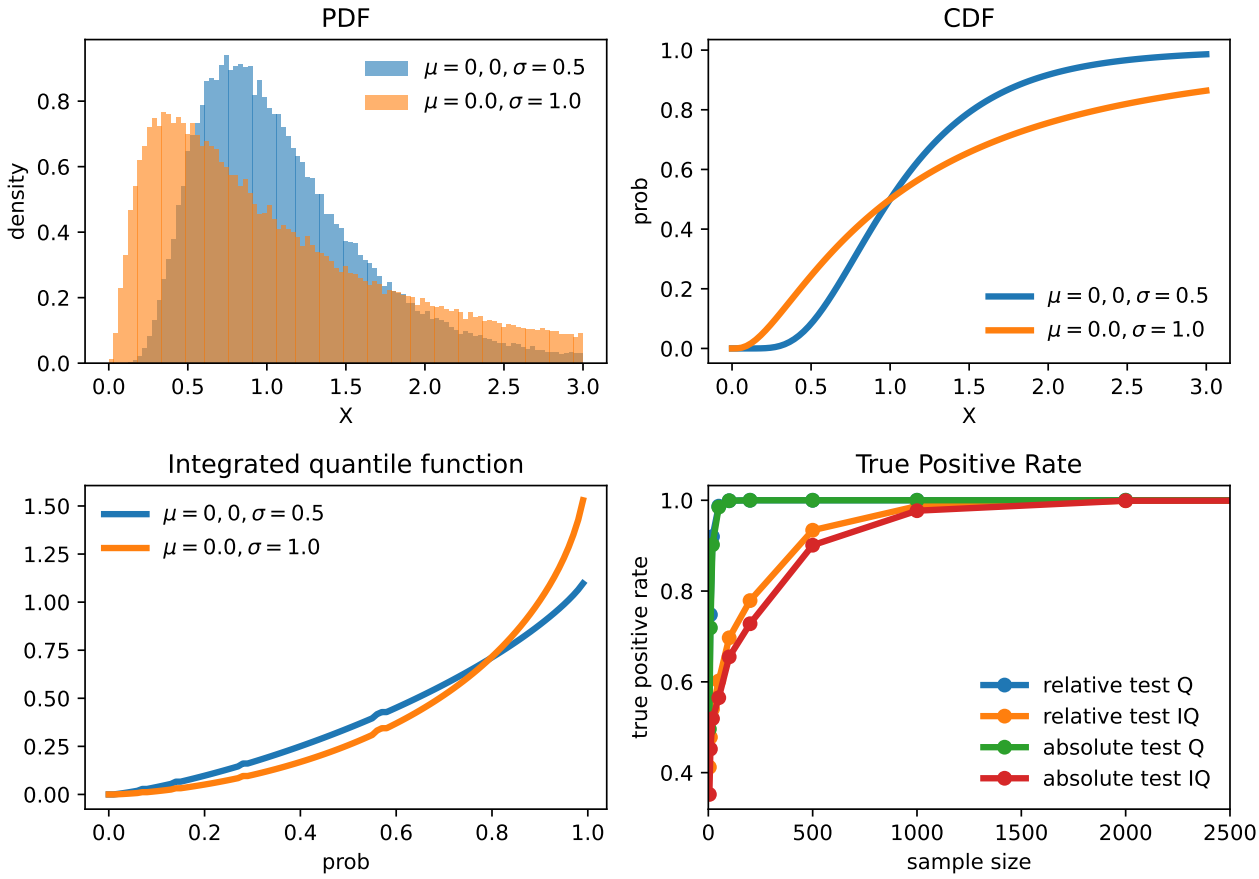


Figure 6: True Positive Rate vs sample size for **Lognormal distributions** generated as $X = e^{\mu+\sigma Z}$, where Z is a standard Gaussian variable. We compute the True Positive Rate of our stochastic dominance as in Fig. 5, but in this case we examine True Positive Rate for heavy-tailed distributions examplified by Lognormal distributions.

L.2. Mix-Instruct

Results for the Mix-Instruct data are shown in Figures 7 and 8, as well as Table 5.

L.3. Toxicity

Toxicity results are in Table 6.

L.4. Ablation Study on toxicity Independent Versus Empirical Copula Portfolio

When Comparing EC and IC portfolio aggregation using R-SSD to rank the LLM we see in Figures 9 and 10 that the two aggregation approaches lead to same ranking. While IC computational complexity is linear in the number of points , EC is quadratic. Given the correspondence in ranking IC is a more efficient aggregation technique.

L.5. Fat Left Tails of Metrics and Inconsistency of Mean-Variance with SSD

When metrics evaluated have fat tails, the Mean-Variance ranking can be inconsistent with the SSD. See Table 7.

	Open assistant	koala	alpaca	llama 7b	flan-t5	stablelm	Vicuna	Dolly (v2)	Moss 6b	ChatGLM	mpt-7b instruct	mpt-7b
Mean Win Rates												
RA(MWR @ M)	**1**	6	**2**	8	5	7	**3**	10	9	4	11	12
MWR @P(IC)	**1**	5	**2**	7	6	8	**3**	9	10	4	11	12
Relative FSD												
RA(R-FSD @ M)	**1**	6	**2**	5	8	11	4	10	7	**3**	9	12
R-FSD @P(IC)	**1**	6	**2**	5	11	10	4	8	7	**3**	9	12
R-FSD @ChatGPT	**1**	7	**3**	4	12	11	**2**	8	5	6	9	10
Relative SSD												
RA(R-SSD @ M)	**1**	7	**2**	5	12	10	4	9	6	**3**	8	11
R-SSD @P(IC)	**1**	6	**3**	5	12	11	4	7	8	**2**	9	10
R-SSD @ChatGPT	**1**	8	**3**	4	11	12	**2**	7	5	6	9	10
Mean-Risk Models												
RA($\mu_X - \Gamma_X$) @ M	**1**	7	**2**	5	12	11	4	9	6	**3**	8	10
RA($\mu_X - r_X$) @P(IC)	**1**	6	**3**	5	12	11	4	7	8	**2**	9	10

Table 4: Rankings of models on following instructions according to all tests, with the top 3 ranks highlighted. We see that SSD and Mean – Risk models are consistent. Note that RA($\mu_X - r_X$) @P(IC) denotes the aggregation of rankings produced by ($\mu_X - r_X$) @P(IC) for each r_X in Table 3.

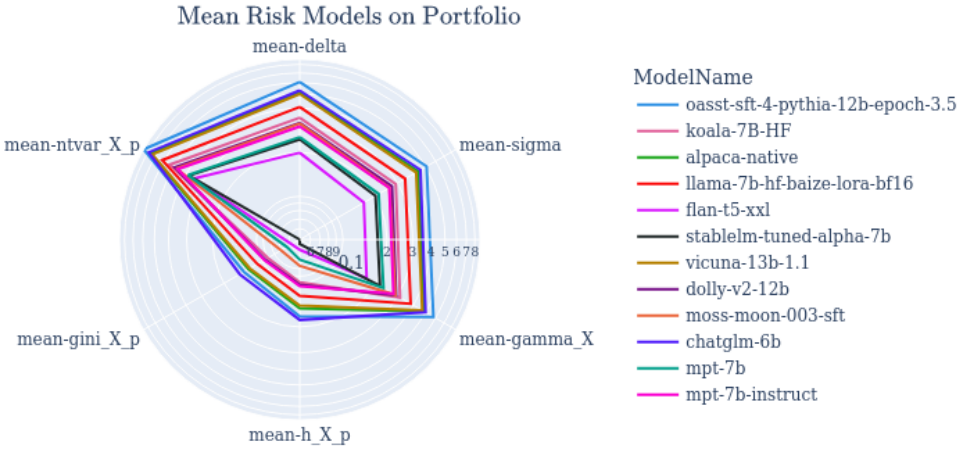


Figure 7: Radar plot of mean – risk models of the portfolio on Mix-Instruct data. Note that the outer models are indeed the ones preferred by SSD in Table 5.

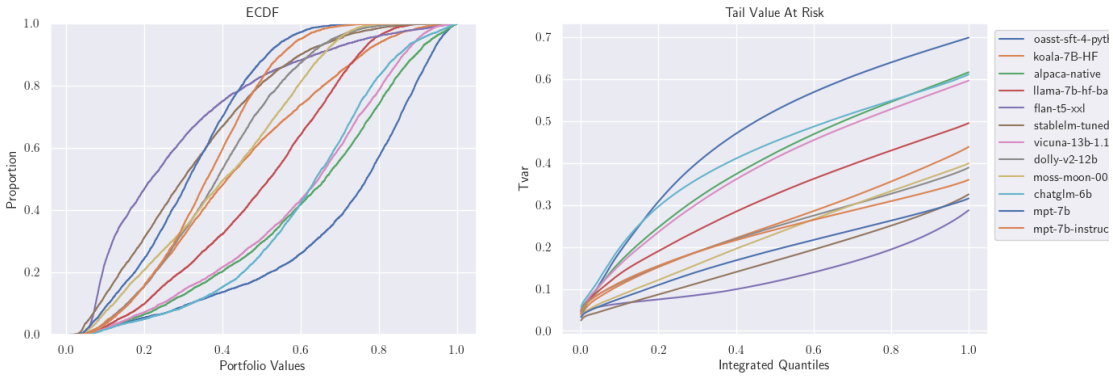


Figure 8: Empirical CDF and TvaR for portfolio on Mix-Instruct data

IC Copula Portfolio

Prompt + gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	llama-2-7b
4	tiiuae-falcon-40b
5	llama-2-70b

gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	llama-2-7b
4	tiiuae-falcon-40b
5	llama-2-70b

EC Copula Portfolio

Prompt + gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	llama-2-7b
4	tiiuae-falcon-40b
5	llama-2-70b

gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	llama-2-7b
4	tiiuae-falcon-40b
5	llama-2-70b

40 K samples

Figure 9: IC versus EC Portfolio Aggregation on Toxicity. Ranking of models using 40 K samples, with independent and Empirical Copula portfolio with R-SSD. We see that the two aggregation methods lead to similar results.

	Open assistant	koala	alpaca	llama 7b	flan-t5	stablelm	Vicuna	Dolly (v2)	Moss 6b	ChatGLM	mpt-7b instruct	mpt-7b
Mean Win Rates												
RA(MWR @ M)	**1**	6	**2**	8	5	7	**3**	10	9	4	11	12
MWR @P(IC)	**1**	5	**2**	7	6	8	**3**	9	10	4	11	12
Relative FSD												
RA(R-FSD @ M)	**1**	6	**2**	5	8	11	4	10	7	**3**	9	12
R-FSD @P(IC)	**1**	6	**2**	5	11	10	4	8	7	**3**	9	12
Relative SSD												
RA(R-SSD @ M)	**1**	7	**2**	5	12	10	4	9	6	**3**	8	11
R-SSD @P(IC)	**1**	6	**3**	5	12	11	4	7	8	**2**	9	10
R-SSD @ChatGPT	**1**	8	**3**	4	11	12	**2**	7	5	6	9	10
Absolute FSD												
ε-FSD @P(IC) ε=0.08	**1**	6	**2**	5	10	11	4	7	8	**3**	9	12
ε-FSD @P(IC) ε=0.25	**1**	6	**2**	5	12	10	4	7	8	**3**	9	11
ε-FSD @P(IC) ε=0.4	**1**	6	**2**	5	12	10	4	8	7	**3**	9	11
Absolute SSD												
ε-SSD @P(IC) $\varepsilon = 0.08$	**1**	6	**3**	5	12	11	4	7	8	**2**	9	10
ε-SSD @P(IC) $\varepsilon = 0.25$	**1**	6	**3**	5	12	11	4	8	7	**2**	9	10
ε-SSD @P(IC) ε=0.4	**1**	6	**3**	5	12	11	4	7	8	**2**	9	10
Mean-Risk Models												
RA($\mu_X - r_X$) @P(IC)	**1**	6	**3**	5	12	11	4	7	8	**2**	9	10

Table 5: Mix instruct Extended Results.

Scenario	Llama 2 7b	Llama 2 13b	Llama 2 70b	MosaicML MPT 30b	Tiiuae Falcon 40b
Toxic Prompts					
RA(R-FSD @M) (Gen Only)	**3**	**2**	4	**1**	5
R-FSD @P(IC)(IC)(Gen Only)	**2**	**3**	4	**1**	5
RA(R-SSD @M) (Gen Only)	**3**	**2**	4	**1**	5
R-SSD@P(IC)(IC) (Gen Only)	**3**	**2**	4	**1**	5
RA(R-FSD @M) (Prompt + Gen)	**2**	**3**	**1**	4	5
R-FSD @P(IC)(IC)(Prompt + Gen)	**2**	**3**	**1**	4	5
RA(R-SSD @M) (Prompt + Gen)	**2**	**3**	**1**	4	5
R-SSD @P(IC)(IC) (Prompt + Gen)	**2**	**3**	**1**	4	5
Non-Toxic Prompts					
RA(R-FSD @M) (Gen Only)	**1**	**2**	4	**3**	5
R-FSD @P(IC)(IC) (Gen Only)	**1**	**2**	**3**	4	5
RA(R-SSD @M) (Gen Only)	**1**	**2**	**3**	4	5
R-SSD @P(IC)(IC) (Gen Only)	**1**	**2**	**3**	4	5
RA(R-FSD @M) (Prompt + Gen)	**3**	**2**	4	**1**	5
R-FSD @P(IC) (Prompt + Gen)	**1**	**2**	4	**3**	5
RA(R-SSD @M) (Prompt + Gen)	**1**	**2**	**3**	4	5
R-SSD @P(IC) (Prompt + Gen)	**1**	**2**	4	**3**	5
All Combined (Toxic + Non-Toxic Prompts)					
RA(R-FSD @M) (Gen Only)	**2**	**3**	5	**1**	4
R-FSD @P(IC) (Gen Only)	**2**	**3**	5	**1**	4
RA(R-SSD @M) (Gen Only)	**2**	**3**	5	**1**	4
R-SSD @P(IC) (Gen Only)	**2**	**3**	5	**1**	4
RA(R-FSD @M) (Prompt + Gen)	**3**	4	5	**1**	**2**
RA(R-FSD @M) (Prompt + Gen)	**3**	4	5	**1**	**2**
R-SSD @P(IC) (Prompt + Gen)	**3**	4	5	**1**	**2**
R-SSD @P(IC) (Prompt + Gen)	**3**	4	5	**1**	**2**

Table 6: Toxicity Ranking Extended Results

IC Copula Portfolio

Prompt + gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	tiiuae-falcon-40b
4	llama-2-7b
5	llama-2-70b

gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	tiiuae-falcon-40b
4	llama-2-7b
5	llama-2-70b

EC Copula Portfolio

Prompt + gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	tiiuae-falcon-40b
4	llama-2-7b
5	llama-2-70b

gen

	SSD
1	mosaicml-mpt-30b
2	llama-2-13b
3	tiiuae-falcon-40b
4	llama-2-7b
5	llama-2-70b

20 K samples

Figure 10: IC versus EC Portfolio Aggregation on Toxicity. Ranking of models using 20 K samples, with independent and Empirical Copula portfolio with R-SSD. We see that the two aggregation methods lead to similar results.

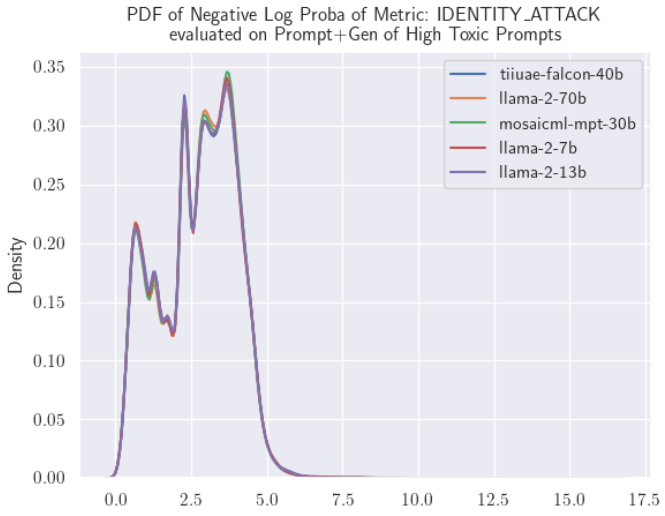


Figure 11: Identity Attack Metric distribution computed on Prompt+Generation output of Highly Toxic Prompts

Scenario	Llama 2 7b	Llama 2 13b	Llama 2 70b	MosaicML MPT 30b	Tiiuae Falcon 40b
Non Toxic Prompts					
Identity Attack Metric Gen evaluation					
Mean - Sigma	**1**	**3**	4	**2**	5
Mean - Gamma	**2**	**3**	4	**1**	5
Mean - nTvAR	**2**	**3**	4	**1**	5
SSD	**2**	**3**	4	**1**	5
Threat Metric Prompt + Gen evaluation					
Mean - Sigma	**1**	**3**	**2**	4	5
Mean - Gamma	**1**	**2**	**3**	5	4
Mean - nTvAR	**1**	**2**	**3**	5	4
SSD	**1**	**2**	**3**	5	4

Table 7: Inconsistency of Mean - Sigma on Toxicity Metrics with SSD and other mean-risk models. This is a due to the fact the metric evaluated may a have a fat left tail see Figures 11 and 13.

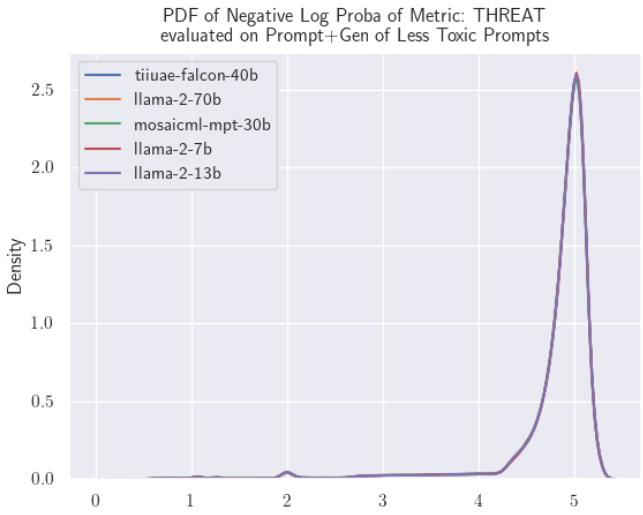


Figure 12: Threat Metric distribution computed on Prompt+Generation output of Less Toxic Prompts

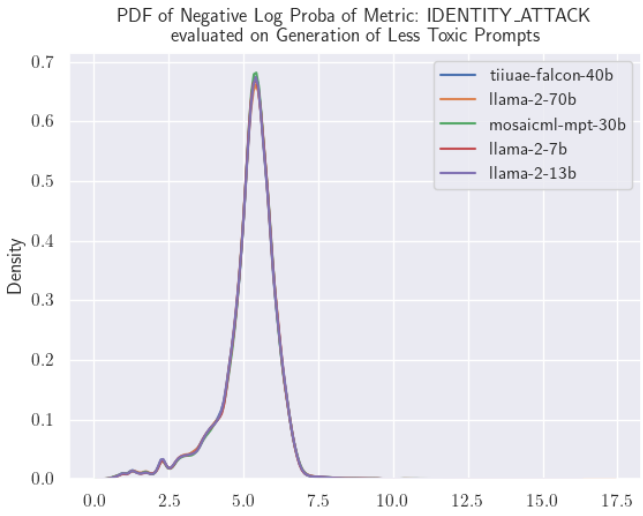


Figure 13: Identity Attack Metric distribution computed on Generation output of Less Toxic Prompts