Risk Assessment and Statistical Significance in the Age of Foundational Models

IBM **Research**

Apoorva Nitsure

Youssef Mroueh, Mattia Rigotti, Kristjan Greenewald, Brian Belgodere, Mikhail Yurochkin, Jiri Navratil, Igor Melnyk, and Jerret Ross

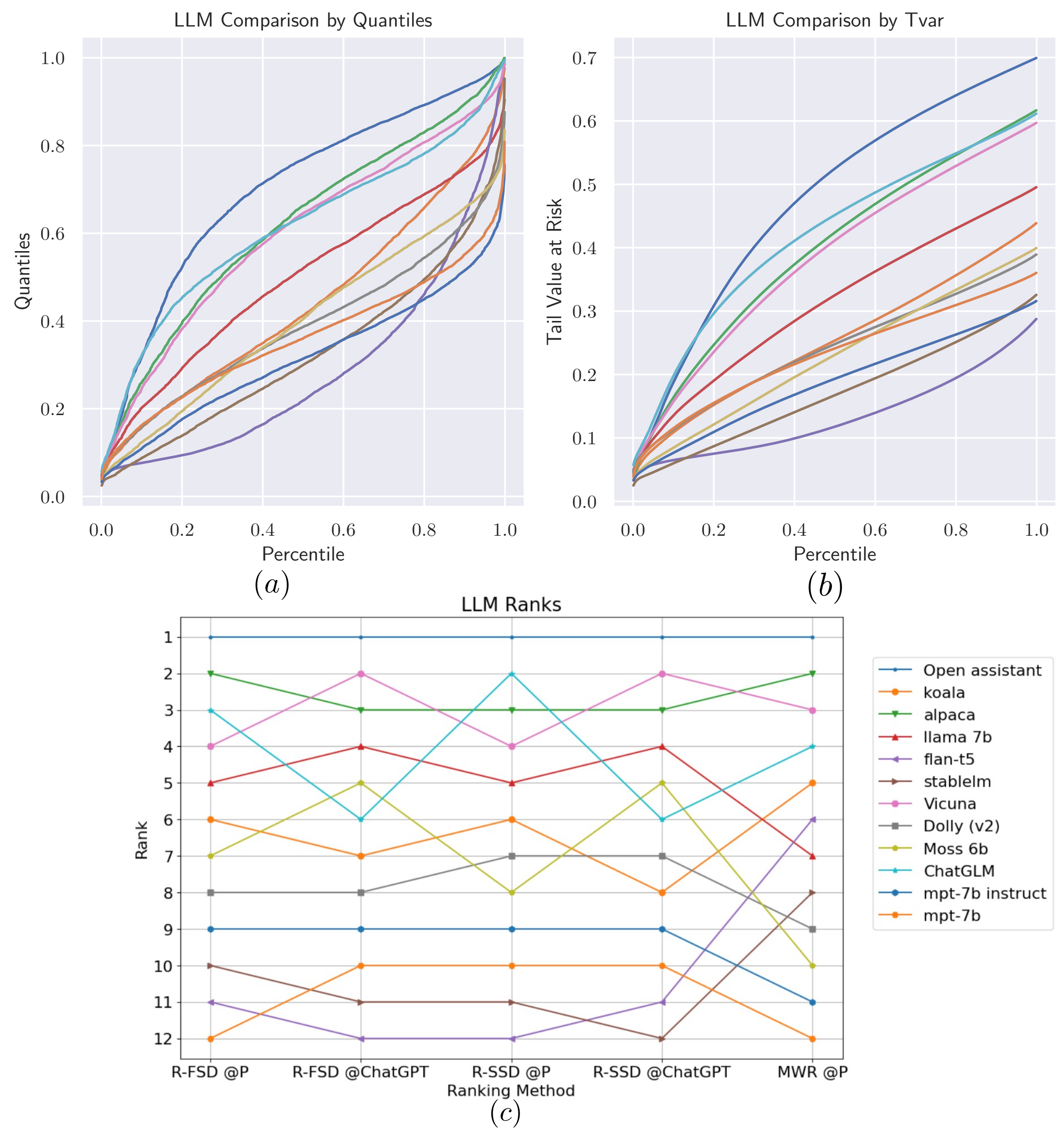
A distributional framework for holistic evaluation and comparison of multiple LLMs across varied metrics with quantified statistical significance

Relative testing based approach using first and second stochastic orders for ranking

Borrows mathematical finance and econometrics concepts to create a metrics portfolio for each model that aggregates multiple metrics for a holistic comparison

Motivation

- Metrics assess socio-technical risks of LLMs such toxicity, factuality and so on (high value of metrics correspond to low risk)
- A risk averse user prefers models that not only perform well on average but most importantly do not exhibit risky tail profiles
- Mean Win Rate does not take into account the risk profile



(a) **First order evaluation (FSD):** LLM comparison based on Quantile of metrics (b) **Second order evaluation (SSD):** LLM comparison based on Tvar (Tail Value at Risk/Integrated quantiles). Tvar teases apart Risky models. (c) **Validation** of automatic metrics with FSD and SSD w.r.t to chatGPT score and Mean Win Rates.

Stochastic Orders: Comparing Distributions

First Order Stochastic Dominance

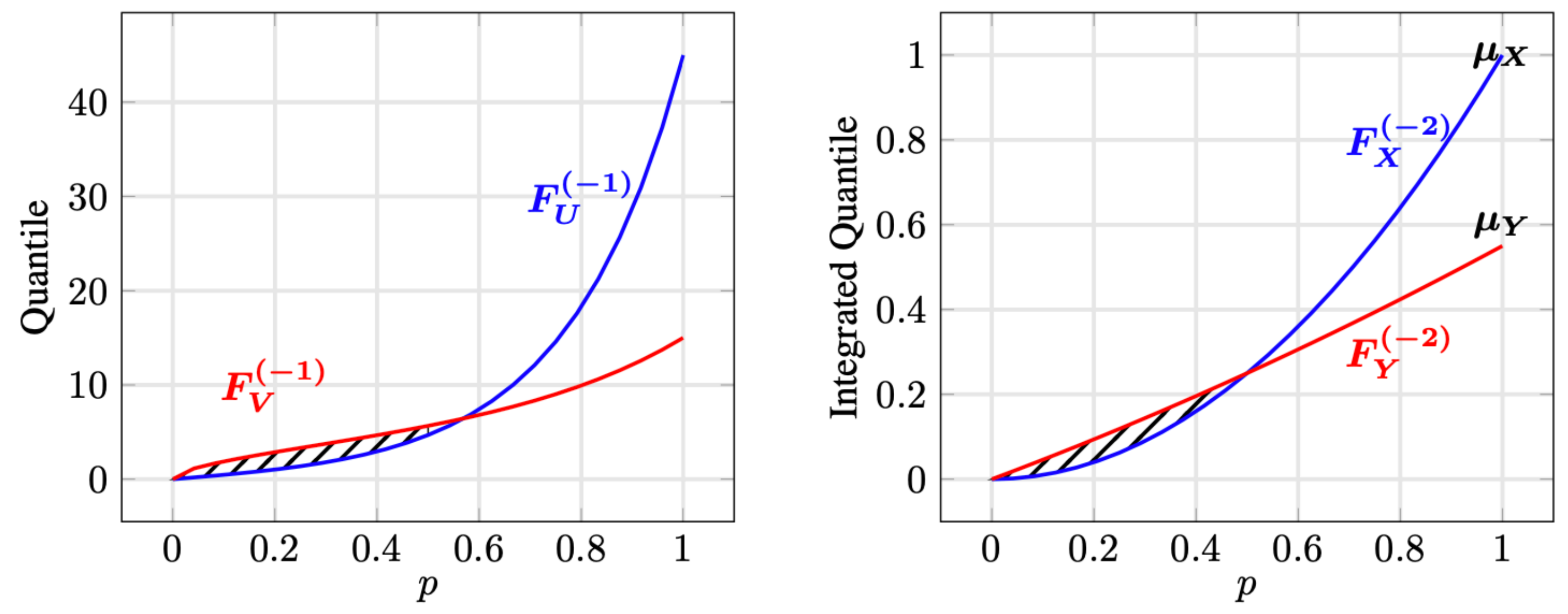
$$X \underset{\text{FSD}}{\succcurlyeq} Y \iff F_X^{(-1)}(p) \geq F_Y^{(-1)}(p), \forall p \in (0,1]$$

Second Order Stochastic Dominance

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

X, Y are real-valued random variables & use the right-continuous cumulative distribution (CDF) as a performance function. $F_X^{(-1)}, F_X^{(-2)}$ are the first and second quantile functions

Relaxations Based on Violation Ratios



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

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

$$X \underset{\varepsilon-FSD}{\succcurlyeq} Y \iff \varepsilon_{\mathrm{W}_2}(F_X, F_Y) = \frac{\int_0^1 (F_Y^{(-1)}(t) - F_X^{(-1)}(t))_+^2 dt}{\mathrm{W}_2^2(F_X, F_Y)} \leq \varepsilon$$

$$X \underset{\varepsilon-SSD}{\succcurlyeq} Y \iff \varepsilon_{IQ}(F_X, F_Y) = \frac{\int_0^1 (F_Y^{(-2)}(t) - F_X^{(-2)}(t))_+^2 dt}{d_{IQ}^2(F_X, F_Y)} \leq \varepsilon$$

For a confidence level α, test using a CLT and Bootstrapping:

$$\varepsilon(F_X^n, F_Y^m) \leq \varepsilon_0 + \sqrt{\frac{m+n}{mn}} \sigma^2(F_X, F_Y) \Phi^{-1}(\alpha)$$

Relative Testing:

Given pairwise violation ratios of N models, compute a one versus all violation ratio for each model (OVR) and compare OVRs of each pair of models. This test does not need a threshold.

From Pairwise Testing to getting a Rank: Use the Borda Algorithm to obtain a rank from pairwise testings

Metrics Aggregation via Portfolio:

Combine N metrics by taking the geometric mean of their CDFS

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

Apply stochastic dominance to resulting portfolio Aggregation

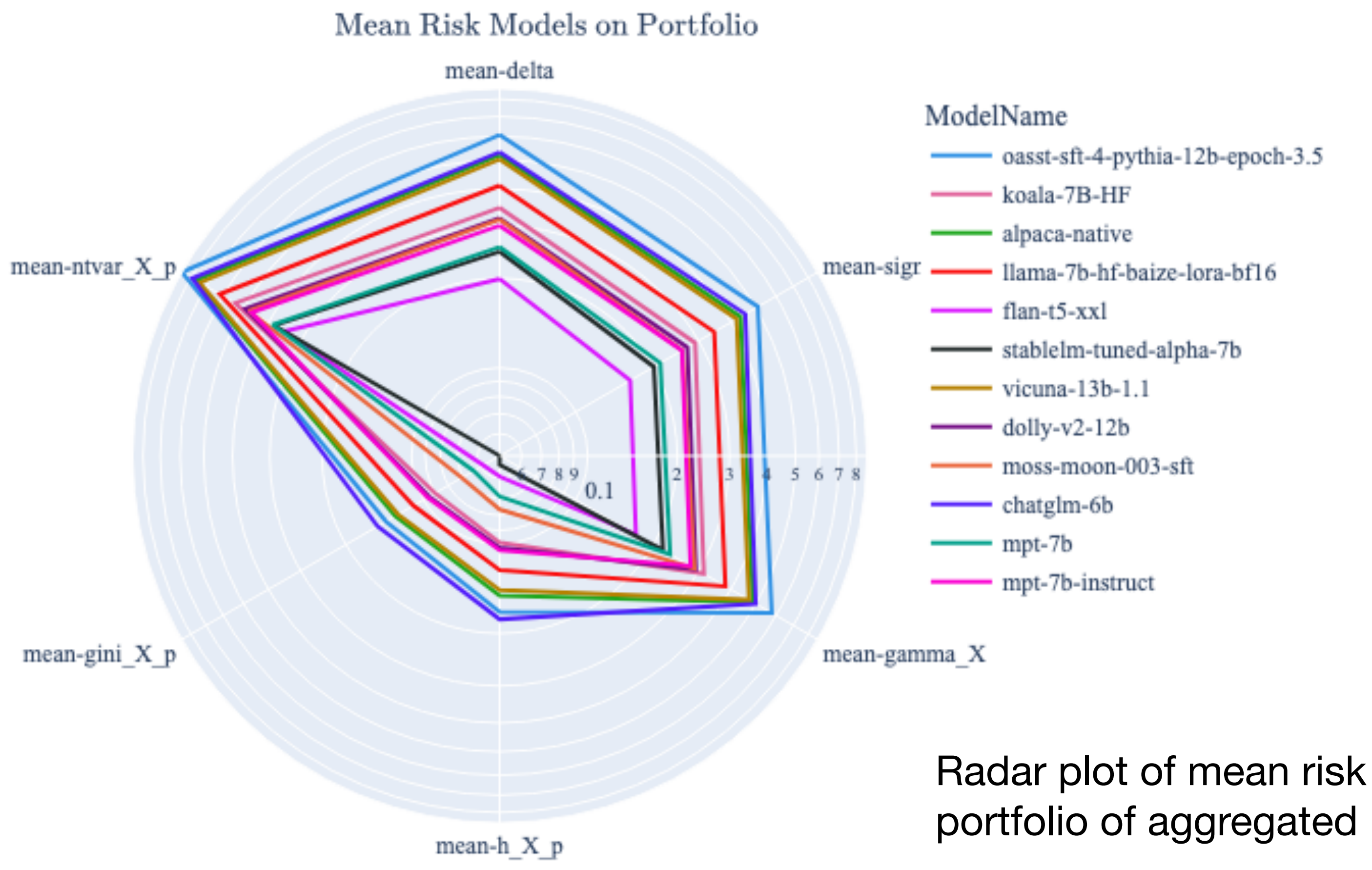
Test Case: Mix Instruct Dataset

	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	**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	**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	**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	**1**	6	**3**	5	12	11	4	7	8	**2**	9	10

The table shows ranking of different LLMs based on instruction following evaluation metrics obtained using our framework. We observe our method helps a user do a risk based assessment while choosing models as SSD based comparison aligns with 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

Risk Models and their α-consistency with SSD



Radar plot of mean risk model scores on portfolio of aggregated metrics