Information Theoretic Guarantees For Policy Alignment In Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Policy alignment of large language models refers to constrained policy optimization, where the policy is optimized to maximize a reward while staying close to a reference policy based on an f-divergence like KL divergence. The best of n alignment policy selects the sample with the highest reward from n independent samples. Recent work shows that the reward improvement of the aligned policy scales as $\sqrt{\mathsf{KL}}$, with an explicit bound on the KL for best of n policies. We show that this $\sqrt{\mathsf{KL}}$ bound holds if the reference policy's reward has sub-gaussian tails. For best of n policies, the KL bound applies to any f-divergence through a reduction to exponential order statistics using the Rényi representation. Tighter control can be achieved with Rényi divergence if additional tail information is known. Finally, we demonstrate how these bounds transfer to golden rewards, resulting in decreased golden reward improvement due to proxy reward overestimation and approximation errors.

1 Introduction

Aligning Large Language Models (LLMs) with human preferences allows a tradeoff between maintaining the utility of the pre-trained reference model and the alignment of the model with human values such as safety or other socio-technical considerations. Alignment is becoming a crucial step in LLMs training pipeline, especially as these models are leveraged in decision making as well as becoming more and more accessible to the general public. Policy alignment starts by learning a reward model that predicts human preferences, these reward models are typically fine-tuned LLMs that are trained on pairwise human preference data (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). The reward is then optimized using *training time alignment* i.e via policy gradient based reinforcement learning leading to the so called *Reinforcemnent Learning from Human Feedback* (RLHF) (Christiano et al., 2017). RLHF ensures that the reward is maximized while the policy π stays close to the initial reference policy π_{ref} in the sense of the Kullback-Leibler divergence $\mathsf{KL}(\pi||\pi_{\text{ref}})$. Other variants of these training time alignment have been proposed via direct preference optimization (Rafailov et al., 2024) (Zhao et al., 2023) (Ethayarajh et al., 2024). Another important paradigm for optimizing the reward is *test time alignment* via best of n sampling from the reference policy and retaining the sample that maximizes the reward. The resulting policy is known as the *best of n policy*. The best of n policy is also used in controlled decoding settings (Yang & Klein, 2021; Mudgal et al., 2023) and in fine-tuning LLMs to match the best of n policy responses (Touvron et al., 2023).

(Gao et al., 2023) and (Hilton & Gao, 2022) studied the scaling laws of reward models optimization in both the RL and the best of n setups. (Gao et al., 2023) distinguished between "golden reward" that can be thought of as the golden human preference and "proxy reward" which is trained to predict the golden reward. For proxy rewards (Gao et al., 2023) found experimentally for both RL and best of n policies that the reward improvement on the reference policy scales as $\sqrt{\mathsf{KL}(\pi||\pi_{\text{ref}})}$. Similar observations for reward improvement scaling in RL were made in (Bai et al., 2022). For golden rewards, (Gao et al., 2023) showed for both RL and best of n policies that LLMs that optimize the proxy reward suffer from over-optimization in the sense that as the policy drifts from the reference policy, optimizing the proxy reward results in deterioration of the golden reward.

This phenomena is referred to in (Gao et al., 2023) (Hilton & Gao, 2022) as Goodhart's law. A qualitative plot of scaling laws discovered in (Gao et al., 2023) is given in Figure 1. For the best of n policy, most works in this space assumed that $\mathsf{KL}(\pi||\pi_{\text{ref}}) = \log(n) - \frac{n-1}{n}$ (Stiennon et al., 2020; Coste et al., 2024; Nakano et al., 2021; Go et al., 2024; Gao et al., 2023). Recently Beirami et al. (2024) showed that this is in fact an inequality under the assumption that the reward is one to one map (a bijection) and for finite alphabets.

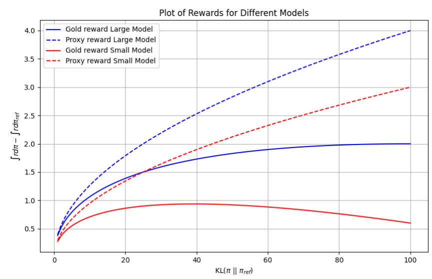

Figure 1: Qualitiative plot of centered rewards vs. KL of Proxy and Gold Rewards for both Best of n and RL policies. (See Fig. 1 a) and b) in (Gao et al., 2023) for scaling laws in policy alignment).

The main contributions of this paper are :

1. In Theorem 1 (Section 2), we provide a new proof for the best of n policy inequality $\mathsf{KL}(\pi||\pi_{\text{ref}}) \leq \log(n) - \frac{n-1}{n}$, showing it results from the data processing inequality of KL. We extend this beyond (Beirami et al., 2024)'s setup of one-to-one rewards and finite alphabets to surjective rewards and beyond finite alphabets, also giving conditions for equality and generalizing to f-divergences and Rényi divergences.

2. In Section 3, we show that the policy improvement scaling laws in (Gao et al., 2023) are information-theoretic upper bounds, derived from transportation inequalities with KL under sub-gaussian reward tails. We discuss how the KL dependence is driven solely by the reward tails and can only improve if they are fatter than sub-gaussian (e.g., sub-gamma or sub-exponential).

3. In Theorem 4, we examine the tightness of these bounds when the optimized policy's tails are known, deriving new transportation inequalities for Rényi divergence D_α for $\alpha \in (0, 1)$. We show that the $\sqrt{\mathsf{KL}}$ upper bound cannot be met, reflecting Goodhart's law (Gao et al., 2023).

4. In Section 4 , we study the transfer of transportation inequalities from proxy to golden rewards, proving that golden reward improvement is limited by the proxy reward's overestimation, as reported empirically in (Gao et al., 2023).

Our work answers positively the following question: *"Can we before doing any alignment predict the upper bound of any alignment from looking only on the histogram of the reward of the reference models?"* Indeed from the tails of the reward under the reference model we can predict the upper bound for alignment as shown in Figures 2 and 3.

2 The Alignment Problem

2.1 RLHF: A Constrained Policy Optimization Problem

Let $\mathcal{X}$ be the space of prompts and $\mathcal{Y}$ be the space of responses $y \in \mathcal{Y}$ from a LLM conditioned on a prompt $x \in \mathcal{X}$. The reference LLM is represented as policy $\pi_{\text{ref}}(y|x)$, i.e as a conditional probability on $\mathcal{Y}$ given a prompt $x \in \mathcal{X}$. Let $\rho_{\mathcal{X}}$ be a distribution on prompts, and a r a reward, $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, r represents a safety or alignment objective that is desirable to maximize.
Given a reference policy π_{ref}, the goal of alignment is to find a policy π^* that maximizes the reward r and

that it is still close to the original reference policy for some positive $\Delta > 0$:

$$\pi^*_{y|x} = \arg\max_{\pi_{y|x}} \mathbb{E}_{x\sim\rho_{\mathcal{X}}} \mathbb{E}_{y\sim\pi(.|x)} r(x,y)$$
$$\text{s.t} \int_{\mathcal{X}} \mathsf{KL}(\pi(y|x)||\pi_{\text{ref}}(y|x)) d\rho_{\mathcal{X}}(x) \leq \Delta, \tag{1}$$

where $\mathsf{KL}(\pi(y|x)||\pi_{\text{ref}}(y|x)) = \mathbb{E}_{y\sim\pi.|x} \log\left(\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}\right)$. With some abuse of notation, we write $\pi(x,y) = \pi(y|x)\rho_{\mathcal{X}}(x)$ and $\pi_{\text{ref}}(x,y) = \pi_{\text{ref}}(y|x)\rho_{\mathcal{X}}(x)$. Let $\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ be joint probability defined on $\mathcal{X}\times\mathcal{Y}$ that has $\rho_{\mathcal{X}}$ as marginal on $\mathcal{X}$. Hence we can write the alignment problem (1) in a more compact way as follows:

$$\sup_{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \int r d\pi \text{ s.t } \mathsf{KL}(\pi||\pi_{\text{ref}}) \leq \Delta. \tag{2}$$

For $\beta > 0$, we can also write a penalized form of this constrained policy optimization problem as follows:

$$\sup_{\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})} \int r d\pi - \frac{1}{\beta}\mathsf{KL}(\pi||\pi_{\text{ref}}).$$

It is easy to see that the optimal policy of the penalized problem is given by:

$$\pi_{\beta,r}(y|x) = \frac{\exp(\beta r(x,y))\pi_{\text{ref}}(y|x)}{\int \exp(\beta r(x,y)) d\pi_{\text{ref}}(y|x)}, \rho_{\mathcal{X}} \text{almost surely.} \tag{3}$$

The constrained problem (2) has a similar solution (See for e.g (Yang et al., 2024)):

$$\pi_{\lambda_\Delta,r}(y|x) = \frac{\exp(\frac{r(x,y)}{\lambda_\Delta})\pi_{\text{ref}}(y|x)}{\int \exp(\frac{r(x,y)}{\lambda_\Delta}) d\pi_{\text{ref}}(y|x)}, \rho_{\mathcal{X}} \text{almost surely,} \tag{4}$$

where $\lambda_\Delta > 0$ is a lagrangian that satisfies $\int_{\mathcal{X}} \mathsf{KL}(\pi_{\lambda_\Delta,r}(y|x)||\pi_{\text{ref}}(y|x)) d\rho_{\mathcal{X}}(x) = \Delta$.

2.2 Best of n Policy Alignment

Let X be the random variable associated with prompts such that $\text{Law}(X) = \rho_{\mathcal{X}}$. Let Y be the random variable associated with the conditional response of π_{ref} given X. Define the conditional reward of the reference policy :

$$R(Y)|X := r(X,Y) \text{ where Y } \sim \pi_{\text{ref}}(.|X),$$

we assume that $R(Y)|X$ admits a CDF denoted as $F_{R(Y)|X}$ and let $F^{-1}_{R(Y)|X}$ be its quantile:

$$F^{(-1)}_{R(Y)|X}(p) = \inf\{\eta : F_{R(Y)|X}(\eta) \geq p\} \text{ for } p \in [0,1].$$

Let $Y_1 \dots Y_n$ be independent samples from $\pi_{\text{ref}}(.|X)$. We define the best of n reward as follows:

$$R^{(n)}(Y)|X = \max_{i=1\dots n} R(Y_i)|X, \tag{5}$$

this the maximum of n iid random variables with a common CDF $F_{R(Y)|X}$. The best of n policy corresponds to $Y^{(n)}|X := \arg\max_{i=1\dots n} r(X,Y_i)$. We note $\pi^{(n)}_{r,\text{ref}}(.|X)$ the law of $Y^{(n)}|X$. $\pi^{(n)}_{r,\text{ref}}$ is referred to as the best of n alignment policy. We consider two setups for the reward:

Assumption 1 *We assume that the reward r is a one to one map for a fixed x, and admits an inverse $h_x : \mathbb{R} \to \mathcal{Y}$ such that $h_x(r(x,y)) = y$.*

This assumption was considered in (Beirami et al., 2024). Nevertheless this assumption is strong and not usually meet in practice, we weaken this assumption to the following:

Assumption 2 *We assume that there is a stochastic map H_X such that $H_X(R_{Y|X})) \stackrel{d}{=} Y|X$ and $H_X(R_{Y^{(n)}|X})) \stackrel{d}{=} Y^{(n)}|X$.*

Under Assumption 2, the reward can be surjective which is more realistic but we assume that there is a stochastic map that ensures invertibility not point-wise but on a distribution level. Our assumption means that we have conditionnaly on X: $R|X \to Y|X$ form a markov chain i.e exists $A(Y|R, X)$ so that $P_{Y|X} = A(Y|R, X)P_{R|X}$, and $P_{Y^{(n)}|X} = A(Y|R, X)P_{R^{(n)}|X}$. *Note that Assumption 2, will always hold from Bayes rule as long as this Markov kernel A is well defined.*

Best of n Policy KL Guarantees: A reduction to Exponentials. In what follows for random variables Z, Z' with laws $p_Z, p_{Z'}$ we write interchangeably: $\mathsf{KL}(p_Z||p_{Z'}) = \mathsf{KL}(Z||Z')$. Let us start by looking at $\mathsf{KL}\left[R^{(n)}(Y)||R(Y)\Big|X\right]$ the KL divergence between the conditional reward of the best of n policy and that of the reference policy. Let $E \sim Exp(1)$, the optimal transport map $F^{-1}_{R(Y)|X} \circ F_E$ from the exponential distribution E to $R(Y)|X$ (See for example Theorem 2.5 in (Santambrogio, 2015): E is atomless, but $R(Y)|X$ can be discrete valued) allows us to write:

$$R(Y)|X \stackrel{d}{=} F^{-1}_{R(Y)|X} \circ F_E(E), \tag{6}$$

where $\stackrel{d}{=}$ means equality in distribution. On the other hand, let $R^{(1)}(Y)|X \leq \cdots \leq R^{(n)}(Y)|X$ be the order statistics of the rewards of n independent samples $Y_i, i = 1 \dots n$, $Y_i \sim \pi_{\text{ref}}(.|X)$. The order statistics refer to sorting the random variable from the minimum (index (1)) to the maximum (index (n)). Consider n independent exponential $E_1, \dots E_n$, where $E_i \sim \exp(1)$, and their order statistics $E^{(1)} \leq E^{(2)} \leq \dots E^{(n)}$. The Rényi representation of order statistics (Rényi, 1953), similar to the Optimal Transport (OT) representation allows us to express the distribution of the order statistics of the rewards in terms of the order statistics of exponentials as follows:

$$\begin{aligned}&\left(R^{(1)}(Y)|X, \dots, R^{(n)}(Y)|X\right)\\ &\stackrel{d}{=} \left(F^{-1}_{R(Y)|X} \circ F_E(E^{(1)}), \dots, F^{-1}_{R(Y)|X} \circ F_E(E^{(n)})\right).\end{aligned} \tag{7}$$

The central idea in the Rényi representation is that the mapping $F^{-1}_{R(Y)|X} \circ F_E$ is monotonic and hence ordering preserving and by the OT representation each component is distributed as $R(Y)|X$. See (Boucheron & Thomas, 2012) for more account on the Rényi representation of order statistics. Note that we could have used uniform random variables instead of exponential, we use exponentials to stay faithful to Rényi representation as exponential order statistics have nice properties.

Hence using the OT representation in (6) and the Rényi representation of the maximum (7), we can reduce the KL between the rewards to a KL on functions of exponentials and their order statistics:

$$\begin{aligned}&\mathsf{KL}\left[R^{(n)}(Y)||R(Y)\Big|X\right]\\ &= \mathsf{KL}\left(F^{-1}_{R(Y)|X} \circ F_E(E^{(n)})\Big|\Big|F^{-1}_{R(Y)|X} \circ F_E(E)\right)\\ &= \mathsf{KL}(T_X(E^{(n)})||T_X(E)),\end{aligned} \tag{8}$$

where $T_X = F^{-1}_{R(Y)|X} \circ F_E = F^{-1}_{(r(X,.))_\sharp \pi_{\text{ref}}(.|X)} \circ F_E$.

Under Assumption 1 we can write samples from the best of n policy as $Y^{(n)}|X = h_X(R^n(Y))|X$ and from the reference policy as $Y|X = h_X(R(Y))|X$. Hence we have by the data processing inequality (DPI) for the KL divergence (See for e.g (Polyanskiy & Wu, 2023)) under Assumption 1:

$$\begin{aligned}&\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}|X) = \mathsf{KL}(Y^{(n)}||Y|X)\\ &= \mathsf{KL}(h_X(R^n(Y))||h_X(R(Y))|X)\\ &= \mathsf{KL}(R^n(Y)||R(Y)|X) \qquad (9)\\ &\text{By Assumption 1 } h_X \text{ is one to one and DPI tight}\\ &= \mathsf{KL}(T_X(E^{(n)})||T_X(E)) \qquad (10)\\ &\text{Rényi and Optimal Transport Representations (Eq (8))} \qquad (11)\end{aligned}$$

From Equations (8) and (11) we see that relating $\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}|X)$ to $\mathsf{KL}(E^{(n)}||E)$ can be done through a data processing inequality that depends on the propreties of the map T_X and the space $\mathcal{Y}$. The following Lemma gives a closed form expression for $\mathsf{KL}(E^{(n)}||E)$:

Lemma 1 (KL Between Exponential and Maximum of Exponentials) *Let $E \sim \exp(1)$, and $E_1, \ldots E_n$ be iid exponentials and $E^{(n)}$ their maximum, we have:*

$$\mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}. \tag{12}$$

The following result establishes the the best n policy KL expression or bounds in terms of n :

Theorem 1 *The best of n policy satisfies under (i) Assumption 1 (reward one to one) and for finite $\mathcal{Y}$ or under (ii) Assumption 2 (existence of stochastic "inverse") :*

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}) \leq \mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}. \tag{13}$$

Under Assumption 1, for infinite $\mathcal{Y}$ and assuming $F_{R(Y|X)}$ is continuous and strictly increasing for all X we have:

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}) = \mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}. \tag{14}$$

Divergence	$f(x)$	Bound on $\text{D}_f(\pi^{(n)}_{r,\text{ref}}\|\|\pi_{\text{ref}})$
KL	$x\log(x)$	$\log(n) - \frac{n-1}{n}$
Chi-squared	$(x-1)^2$	$\frac{(n-1)^2}{2n-1}$
Total Variation	$f(x) = \frac{1}{2}\|x-1\|$	$(\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}$
Hellinger distance	$(1-\sqrt{x})^2$	$2\frac{(1-\sqrt{n})^2}{n+1}$
Forward KL	$-\log(x)$	$n-1-\log(n)$
α Rényi Divergence	NA	$\frac{1}{(\alpha-1)}\log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right)$

Table 1: Best of n policy f-Divergence and α Rényi Divergence Bounds.

Beirami et al. (2024) showed this result under condition (i) which is not a realistic setting and used the finiteness of $\mathcal{Y}$ to provide a direct proof. Our analysis via chaining DPI and using OT and Rényi representations to reduce the problem to exponentials allows us to extend the result to a more realistic setup under condition (ii) i.e the existence of a stochastic "inverse", without any assumption on $\mathcal{Y}$. Furthermore we unveil under which conditions the equality holds that was assumed to hold in previous works (Stiennon et al., 2020) (Coste et al., 2024; Nakano et al., 2021; Go et al., 2024) (Hilton & Gao, 2022) (Gao et al., 2023).

Our approach of reduction to exponentials using Rényi representation of order statistics and data processing inequalities extends to bounding the f- divergence $\text{D}_f(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}})$ as well as the α Rényi divergence. The Rényi divergence for $\alpha \in (0,1) \cup (1,\infty)$ is defined as follows:

$$D_\alpha(P||Q) = \frac{1}{(\alpha-1)}\log\left(\int p^\alpha(x)q^{1-\alpha}(x)dx\right)$$

the limit as $\alpha \to 1$ coincides with KL, i.e: $D_1(P||Q)) = \mathsf{KL}(P||Q)$. These bounds are summarized in Table 1. Full proofs and theorems are in the Appendix.

Best of n-Policy Dominance on the Reference Policy. The following proposition shows that the best of n policy leads to an improved reward on average:

Proposition 1 *$R^{(n)}$ dominates R in the first order dominance that is $R^{(n)}$ dominates R on all quantiles: $Q_{R^{(n)}}(t) \geq Q_R(t), \forall t \in [0,1]$. It follows that*

Best of n Policy and RL Policy The following proposition discusses the sub-optimality of the best of n policy with respect to the alignment RL objective given in (1):

Proposition 2 *Assume a bounded reward in $[-M, M]$. For $\Delta > 0$ and $n = \exp(\Delta)$ the best of n policy $\pi^{(n)}_{r,\text{ref}}$ and the Δ Constrained RL policy $\pi_{\lambda_\Delta,r}$ (given in (4)) satisfy:*

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\lambda_\Delta,r}) \leq \frac{\sqrt{2\pi}M(e^{\frac{2M}{\lambda_\Delta}}-1)}{\lambda_\Delta}\exp(-\frac{\Delta}{2}).$$

A similar asymptotic result appeared in (Yang et al., 2024) for $\Delta \to \infty$, showing as $n \to \infty$, $\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{\lambda_\Delta,r}) \to 0$, we provide here a non asymptotic result for finite n and finite Δ.

3 Reward Improvement Guarantees Through Transportation Inequalities

Notations Let X be a real random variable. The logarithmic moment generating function of X is defined as follows for $\lambda \in \mathbb{R}$: $\psi_X(\lambda) = \log \mathbb{E}_X e^{\lambda(X-\mathbb{E}X)}$. X is said to be sub-Gaussian with variance σ^2 if : $\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$for all $\lambda \in \mathbb{R}$. We denote $\mathsf{SubGauss}(\sigma^2)$ the set of sub-Gaussian random variables with variance σ^2_{ref}. X is said to be sub-Gamma on the right tail with variance factor σ^2 and a scale parameter $c > 0$ if : $\psi_X(\lambda) \leq \frac{\lambda^2\sigma^2}{2(1-c\lambda)}$for every λ such that $0 < \lambda < \frac{1}{c}$. We denote $\mathsf{SubGamma}(\sigma^2, c)$ the set of left and right tailed sub-Gamma random variables. Sub-gamma tails can be thought as an interpolation between sub-Gaussian and sub-exponential tails.

Scaling Laws in Alignment It has been observed empirically (Coste et al., 2024; Nakano et al., 2021; Go et al., 2024; Hilton & Gao, 2022; Gao et al., 2023) that optimal RL policy $\pi_{\lambda_\Delta,r}$ satisfy the following inequality for a constant σ^2_{ref} :

$$\mathbb{E}_{\pi_{\lambda_\Delta,r}} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2 \mathsf{KL}(\pi_{\lambda_\Delta,r}||\pi_{\text{ref}})}.$$

A similar scaling for best of n policy :

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}} r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2\left(\log n - \frac{n-1}{n}\right)},$$

and those bounds are oftentimes tight even when empirically estimated from samples.

The case of Bounded Rewards and Pinsker Inequality This hints that those bounds are information theoretic and independent of the alignment problem. Indeed if the reward was bounded, a simple application of Pinsker inequality gives rise to $\sqrt{\mathsf{KL}}$ scaling. Let TV be the total variation distance, we have: $\mathsf{TV}(\pi, \pi_{\text{ref}}) = \frac{1}{2}\sup_{||r||_\infty \leq 1} \mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{\frac{1}{2}\mathsf{KL}(\pi||\pi_{\text{ref}})}$. Hence we can deduce that for bounded rewards r with norm infinity $||r||_\infty$ that:

$$\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2||r||^2_\infty \mathsf{KL}(\pi||\pi_{\text{ref}})}.$$

Nevertheless this boundedness assumption on the reward is not realistic, since most reward models are unbounded: quoting Lambert et al. (2024b) " implemented by appending a linear layer to predict one logit or removing the final decoding layers and replacing them with a linear layer" and hence the reward is unbounded by construction. We will show in what follows that those scalings laws are tied to the tails of the reward under the reference policy and are instances of transportation inequalities. Note that the reward can be rescaled and transformed to be bounded, nevertheless Pinsker inequality remains loose, as with transportation inequality we aim at replacing the $\|r\|_\infty$, by a second moment or a standard deviation σ, and typically if $\sigma << \|r\|_\infty$ this leads to tigther bounds.

3.1 Transportation Inequalities with KL Divergence

For a policy $\pi \in \mathcal{P}(\mathcal{Y})$ and for a reward function $r : \mathcal{Y} \to \mathbb{R}$, we note $r_\sharp\pi$, the push-forward map of π through r. The reader is referred to Appendix D.1 for background on transportation inequalities and how they are derived from the so-called Donsker-Varadhan variational representation of the KL divergence. The following Proposition hinges on Lemma 4.14 in (Boucheron et al., 2013)):

Proposition 3 (Transportation Inequalities) *The following inequalities hold depending on the tails of* $r_\sharp \pi_{\text{ref}}$*:*

1. *Assume that* $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$*. For any* $\pi \in \mathcal{P}(\mathcal{Y})$ *that is absolutely continuous with respect to* π_{ref}*, and such that* $\mathsf{KL}(\pi || \pi_{\text{ref}}) < \infty$ *then we have:*
$$|\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r| \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi || \pi_{\text{ref}})}.$$
2. *Assume that* $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGamma}(\sigma^2_{\text{ref}}, c)$*. For any* $\pi \in \mathcal{P}(\mathcal{Y})$ *that is absolutely continuous with respect to* π_{ref}*, and such that* $\mathsf{KL}(\pi || \pi_{\text{ref}}) < \infty$ *then we have:*
$$|\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r| \leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi || \pi_{\text{ref}})} + c\mathsf{KL}(\pi || \pi_{\text{ref}})$$

In particular we have the following Corollary:

Corollary 1 (Expected Reward Improvement) *If* $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$ *the following holds for the optimal RL policy* $\pi_{\lambda_{\Delta,r}}$ *and for the best of* n *policy* $\pi^{(n)}_{r,\text{ref}}$*:*

1. *For the optimal RL policy* $\pi_{\lambda_{\Delta,r}}$ *we have:*
$$\begin{aligned} 0 \leq \mathbb{E}_{\pi_{\lambda_{\Delta,r}}} r - \mathbb{E}_{\pi_{\text{ref}}} r &\leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi_{\lambda_{\Delta,r}} || \pi_{\text{ref}})} \\ &\leq \sqrt{2\sigma^2_{\text{ref}} \Delta}. \end{aligned}$$
2. *For the Best of* n *policy* $\pi^{(n)}_{r,\text{ref}}$*, under Assumption 2 we have:*
$$\begin{aligned} 0 \leq \mathbb{E}_{\pi^{(n)}_{r,\text{ref}}} r - \mathbb{E}_{\pi_{\text{ref}}} r &\leq \sqrt{2\sigma^2_{\text{ref}} \mathsf{KL}(\pi^{(n)}_{r,\text{ref}} || \pi_{\text{ref}})} \\ &\leq \sqrt{2\sigma^2_{\text{ref}} \left(\log n - \frac{n-1}{n}\right)} \end{aligned}$$

A similar statement holds under sub-gamma tails of the reward of the reference model. Item (1) in Corollary 1 shows that the $\sqrt{\sigma^2_{\text{ref}} \mathsf{KL}}$ provides an upper bound on the reward improvement of the alignment under subgaussian tails of the reference reward. Under subgaussian tails of the reference, this information theoretic barrier can not be broken with a better algorithm. On way to improve on the $\sqrt{\mathsf{KL}}$ ceiling is by aiming at having a reference model with a reward that has subgamma tails to improve the upper limit to $\sqrt{\sigma^2_{\text{ref}} \mathsf{KL}} + c\mathsf{KL}$, or to subexponential tails to be linear in the KL. Item (2) can be seen as a refinement on the classical $\sqrt{2\sigma^2_{\text{ref}} \log(n)}$ upper bound on the expectation of maximum of subgaussians see for e.g Corollary 2.6 in (Boucheron et al., 2013). If in addition r is positive and for $X = r_\sharp \pi_{\text{ref}} - \mathbb{E}_{\pi_{\text{ref}}} r$ we have for $t > 0$, $\mathbb{P}(X > t) \geq \mathbb{P}(|g| > t)$, where $g \sim \mathcal{N}(0, \sigma^2_\ell)$ (where σ^2_ℓ is a variance) , then we have a matching lower bound for $\pi^{(n)}_{r,\text{ref}}$ that scales with $\sqrt{\sigma^2_\ell \log(n)}$ for sufficiently large n (See (Kamath, 2015)).

We turn now to providing a bound in high probability on the empirical reward improvement of RL. The following Theorem gives high probability bounds for the excess reward when estimated from empirical samples:

Theorem 2 (High Probability Empirical Reward Improvement For RL) *Assume* $r_\sharp \pi_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$*. Let* $\beta > 1$ *and* $t_0 > 0$*. Let* $\pi_{\beta,r}$ *be the optimal policy of the penalized RL problem given in Equation* (3)*. Let* $R_{i,\beta}$ *and* $R_{i,\text{ref}}, i = 1 \dots m$ *be the rewards evaluated at* m *samples from* $\pi_{\beta,r}$ *and*

$\pi_{\rm ref}$. *Assume that the β-Rényi divergence $D_\beta(\pi_{\beta,r}||\pi_{\rm ref})$ and $\mathsf{KL}(\pi_{\beta,r}||\pi_{\rm ref})$ are both finite. The following inequality holds with probability at least $1-e^{-\frac{mt_0^2}{2\sigma^2_{\rm ref}}}-e^{-m(\beta-1)t_0}$:*

$$\frac{1}{m}\sum_{i=1}^m R_{i,\beta}-\frac{1}{m}\sum_{i=1}^m R_{i,\rm ref}\leq \sqrt{2\sigma^2_{\rm ref}\mathsf{KL}(\pi_{\beta,r}||\pi_{\rm ref})}$$
$$..+\frac{D_\beta(\pi_{\beta,r}||\pi_{\rm ref})-\mathsf{KL}(\pi_{\beta,r}||\pi_{\rm ref})}{\beta}+2t_0.$$

Note that in Theorem 2, we did not make any assumptions on the tails of $r_\sharp\pi_{\beta,r}$ and we see that this results in a biased concentration inequality with a non-negative bias $\frac{D_\beta(\pi_{\beta,r}||\pi_{\rm ref})-\mathsf{KL}(\pi_{\beta,r}||\pi_{\rm ref})}{\beta}\geq 0$. For the best of n policy, if the reward was positive and has a folded normal distribution (absolute value of gaussians), (Boucheron & Thomas, 2012) provides concentration bounds, owing to subgamma tails of maximum of absolute Gaussians.

3.2 Tail Adaptive Transportation Inequalities with the Rényi Divergence

An important question on the tightness of the bounds rises from the bounds in Corollary 1. We answer this question by considering additional information on the tails of the reward under the policy π, and we obtain tail adaptive bounds that are eventually tighter than the one in Corollary 1. Our new bounds leverage a variational representation of the Rényi divergence that uses the logarithmic moment generating function of both measures at hand.

Preliminaries for the Rényi Divergence The Donsker-Varadahn representation of KL was crucial in deriving transportation inequalities. In Shayevitz (2011) the following variational form is given for the Rényi divergence in terms of the KL divergence, for all $\alpha\in\mathbb{R}$

$$(1-\alpha)D_\alpha(P||Q)=\inf_R \alpha\mathsf{KL}(R||P)+(1-\alpha)\mathsf{KL}(R||Q) \tag{15}$$

A similar variational form was rediscovered in (Anantharam, 2018). Finally a Donsker-Varadahn-Rényi representation of D_α was given in (Birrell et al., 2021). For all $\alpha\in\mathbb{R}^+,\alpha\neq 0,1$ we have :

$$\frac{1}{\alpha}D_\alpha(P||Q)=\sup_{h\in\mathcal{H}}\frac{1}{\alpha-1}\log\left(\mathbb{E}_P e^{(\alpha-1)h}\right)-\frac{1}{\alpha}\log\left(\mathbb{E}_Q e^{\alpha h}\right), \tag{16}$$

where $\mathcal{H}=\left\{h\Big|\int e^{(\alpha-1)h}dP<\infty,\int e^{\alpha h}dQ<\infty\right\}$. Birrell et al. (2021) presents a direct proof of this formulation without exploring its link to the representation given in (15), we show in what follows an elementary proof via convex conjugacy, the duality relationship between equations (15) and (16).

Theorem 3 *For $0<\alpha<1$ Equations* (15) *and* (16) *are dual of one another. For $\alpha>1$ they are Toland Dual.*

We collect in what follows elementary lemmas that will be instrumental to derive transportation inequalities in terms of the Rényi divergence. Proofs are given in the Appendix.

Lemma 2 *Let $\alpha\in(0,1)\cup(1,\infty)$, and define $\mathcal{H}=\{h|e^{(\alpha-1)(h-\int hdP)}\in L^1(P),e^{(\alpha)(h-\int hdQ)}\in L^1(Q)\}$. We have for all $h\in\mathcal{H}$ and for $\alpha\in(0,1)\cup(1,\infty)$*

$$\int hdP-\int hdQ\leq\frac{1}{\alpha}D_\alpha(P||Q)$$
$$-\frac{1}{\alpha-1}\log\left(\int e^{(\alpha-1)(h-\int hdP)}dP\right)$$
$$+\frac{1}{\alpha}\log\left(\int e^{\alpha(h-\int hdQ)}dQ\right).$$

Lemma 3 *The following limit holds for the Rényi divergence* $\lim_{\alpha\to 0}\frac{1}{\alpha}D_\alpha(P||Q)=\mathsf{KL}(Q||P)$.

Transportation Inequalities with Rényi Divergence. The following theorem shows that when considering the tails of π we can obtain tighter upper bounds using the Rényi divergence that is more tail adaptive:

Theorem 4 (Tail Adaptive Transportation Inequalities) *Let $\alpha \in (0,1)$. Assume $r_{\sharp}\pi \in \mathsf{SubGauss}(\sigma^2_{\pi})$ and $r_{\sharp}\pi_{\mathrm{ref}} \in \mathsf{SubGauss}(\sigma^2_{\mathrm{ref}})$ then we have for all $\alpha \in (0,1)$:*

$$\mathbb{E}_{\pi} r - \mathbb{E}_{\pi_{\mathrm{ref}}} r \leq \sqrt{2((1-\alpha)\sigma^2_{\pi} + \alpha\sigma^2_{\mathrm{ref}})\frac{D_{\alpha}(\pi||\pi_{\mathrm{ref}})}{\alpha}}. \tag{17}$$

In particular if there exits $\alpha \in (0,1)$ such that $D_{\alpha}(\pi||\pi_{\mathrm{ref}}) \leq \frac{\alpha\sigma^2_{\mathrm{ref}}}{(1-\alpha)\sigma^2_{\pi}+\alpha\sigma^2_{\mathrm{ref}}}\mathsf{KL}(\pi||\pi_{\mathrm{ref}})$, then the tail adaptive upper bound given in Equation (17) is tighter than the one provided by the tails of π_{ref} only i.e $\sqrt{\sigma^2_{\mathrm{ref}}\mathsf{KL}(\pi||\pi_{\mathrm{ref}})}$. Note that this is possible because D_{α} is increasing in $\alpha \in (0,1)$ (van Erven & Harremos, 2014), i.e $D_{\alpha}(\pi||\pi_{\mathrm{ref}}) \leq \mathsf{KL}(\pi||\pi_{\mathrm{ref}})$, and $\frac{\alpha\sigma^2_{\mathrm{ref}}}{(1-\alpha)\sigma^2_{\pi}+\alpha\sigma^2_{\mathrm{ref}}} \leq 1$. Note that taking limits $\alpha \to 0$ (applying Lemma 3) and $\alpha \to 1$, and taking the minimum of the upper bounds we obtain:

$$\mathbb{E}_{\pi} r - \mathbb{E}_{\pi_{\mathrm{ref}}} r \leq \sqrt{2\min(\sigma^2_{\pi_{\mathrm{ref}}}\mathsf{KL}(\pi||\pi_{\mathrm{ref}}), \sigma^2_{\pi}\mathsf{KL}(\pi_{\mathrm{ref}}||\pi))},$$

this inequality can be also obtained by applying Proposition 3 twice: on the tails of π and π_{ref} respectively.

Another important implication of Theorem 4, other than tighter than KL upper bound, is that if we were to change the RL alignment problem (1) to be constrained by $D_{\alpha}, \alpha \in (0,1)$ instead of KL, we may end up with a smaller upper limit on the reward improvement. This D_{α} constrained alignment may lead to a policy that under-performs when compared to a policy obtained with the KL constraint. This was indeed observed experimentally in (Wang et al., 2024a) that used constraints with α- divergences for $\alpha \in (0,1)$ (that are related to Rényi divergences) and noticed a degradation in the reward improvement w.r.t policies obtained using KL.

4 Transportation Inequality Transfer From Proxy to Golden Reward

As we saw in the previous sections, the tightness of $\sqrt{\mathsf{KL}(\pi||\pi_{\mathrm{ref}})}$ upper bound in alignment can be due to the tails of the reward of the aligned policy π (Theorem 4) and to the concentration around the mean in finite sample size (Theorem 2). Another important consideration is the mismatch between the golden reward r^* that one desires to maximize that is expensive and difficult to obtain (for example human evaluation) and a proxy reward r that approximates r^*. The proxy reward r is used instead of r^* in RL and in best of n policy. While we may know the tails of the reward r of the reference and aligned model, we don't have access to this information on the golden reward r^*. We show in this section how to transfer transportation inequalities from r to r^* for RL and Best of n policy.

Proposition 4 (r^* Transportation Inequality for RL Policy) *The following inequality holds:*

$$\begin{aligned}
&\mathbb{E}_{\pi_{\beta,r}} r^* - \mathbb{E}_{\pi_{\mathrm{ref}}} r^* \leq \mathbb{E}_{\pi_{\beta,r}} r - \mathbb{E}_{\pi_{\mathrm{ref}}} r \;.. \\
&.. - \frac{1}{\beta}\log\left(\int e^{\beta(r-r^*-\left(\int r d\pi_{\mathrm{ref}} - \int r^* d\pi_{\mathrm{ref}}\right)} d\pi_{\beta,r^*}\right),
\end{aligned}$$

Assume $r_{\sharp}\pi_{\mathrm{ref}} \in \mathsf{SubGauss}(\sigma^2_{\mathrm{ref}})$, and there exists $\delta > 0$ such that: $\frac{1}{\beta}\log\left(\int e^{\beta(r-r^*-\left(\int r d\pi_{\mathrm{ref}} - \int r^* d\pi_{\mathrm{ref}}\right)} d\pi_{\beta,r^*}\right) \geq \delta\mathsf{KL}(\pi_{\beta,r^*}||\pi_{\mathrm{ref}})$, then we have:

$$\mathbb{E}_{\pi_{\beta,r}} r^* - \mathbb{E}_{\pi_{\mathrm{ref}}} r^* \leq \sqrt{2\sigma^2_{\mathrm{ref}}\mathsf{KL}(\pi_{\beta,r}||\pi_{\mathrm{ref}})} - \delta\mathsf{KL}(\pi_{\beta,r^*}||\pi_{\mathrm{ref}}).$$

Note that $\frac{1}{\beta}\log\left(\int e^{\beta(r-r^*-\left(\int rd\pi_{\rm ref}-\int r^*d\pi_{\rm ref}\right)}d\pi_{\beta,r^*}\right)$ is interpreted here as an interpolation between the mean and the maximum of its argument on the support of π_{β,r^*} (Proposition 9 in (Feydy et al., 2018)). Indeed as $\beta\to 0$, this boils down to the mean on $\int(r-r^*)d\pi_{\beta,r^*}-\left(\int rd\pi_{\rm ref}-\int r^*d\pi_{\rm ref}\right)$ and $\beta\to\infty$ this boils down to $\max_{\mathsf{supp}\pi_{\beta,r^*}}\{r-r^*-\left(\int rd\pi_{\rm ref}-\int r^*d\pi_{\rm ref}\right)\}$. Our assumption means that r overestimates r^* and the overestimation is accentuated as we drift from $\pi_{\rm ref}$ on which r was learned. If r overestimates r^*, there exists $\Delta>0$ such that: $r-r^*-\left(\int rd\pi_{\rm ref}-\int r^*d\pi_{\rm ref}\right)\geq\Delta$ By Jensen inequality we have: $\frac{1}{\beta}\log\left(\int e^{\beta(r-r^*-\left(\int rd\pi_{\rm ref}-\int r^*d\pi_{\rm ref}\right))}d\pi_{\beta,r^*}\right)\geq\frac{1}{\beta}\int\beta(r-r^*-\left(\int rd\pi_{\rm ref}-\int r^*d\pi_{\rm ref}\right))d\pi_{\beta,r^*}\geq\Delta$. Hence our assumption is on the overestimation error $\Delta=\delta\mathsf{KL}(\pi_{\beta,r^*}||\pi_{\rm ref})$. This assumption echoes findings in (Gao et al., 2023) that show that the transportation inequalities suffer from overestimation of proxy reward models of the golden reward (See Figure 8 in (Gao et al., 2023)).

Note that in Proposition 4, we are evaluating the golden reward r^* improvement when using the proxy reward optimal policy $\pi_{\beta,r}$. We see that the golden reward of the RL policy inherits the transportation inequality from the proxy one but the improvement of the reward is hindered by possible overestimation of the golden reward by the proxy model. This explains the dip in performance as measured by the golden reward depicted in Figure 1 and reported in (Gao et al., 2023).

Proposition 5 (r^* Transportation Inequality for Best of n Policy) *Let $\varepsilon>0$. Let r be a surrogate reward such that $\|r-r^*\|_{\infty}\leq\varepsilon$ and assume $r_{\sharp}\pi_{\rm ref}\in\mathsf{SubGauss}(\sigma^2_{\rm ref})$ then the best of n policy $\pi^{(n)}_{r,\rm ref}$ satisfies:*

$$\mathbb{E}_{\pi^{(n)}_{r,\rm ref}}(r^*)-\mathbb{E}_{\pi_{\rm ref}}(r^*)\leq\sqrt{2\sigma^2_{\rm ref}\left(\log(n)-\frac{n-1}{n}\right)}+2\varepsilon\left(\left(\frac{1}{n}\right)^{\frac{1}{n-1}}-\left(\frac{1}{n}\right)^{\frac{n}{n-1}}\right).$$

Transportation inequalities transfers for the best of n policy from r to r^* and pays only an additional error term $\|r-r^*\|_{\infty}\mathsf{TV}(\pi^{(n)}_{r,\rm ref}|\pi_{\rm ref})$, an upper bound of this total variation as a function of n is given in Table 1. As mentioned earlier, if we have lower bounds on the tail of the reference reward, then we also have a lower bound on the reward improvement that scales like $C\sqrt{\sigma^2_{\ell}\log(n)}-2\varepsilon\left(\left(\frac{1}{n}\right)^{\frac{1}{n-1}}-\left(\frac{1}{n}\right)^{\frac{n}{n-1}}\right)$. This is in line with empirical findings in (Hilton & Gao, 2022) (Gao et al., 2023) that showed that best of n policy is resilient as the reward model r gets closer to r^*.

Practical Implications. As the proxy reward may overestimate the golden reward, (Wang et al., 2024b) proposed a reward transformation that reduces the tails of the reward distribution. This, in turn, improves the golden reward by avoiding shortcuts that exploit the reward and target the unreliable tails of the distribution, which overestimate the golden reward with high values. In our theory, this would reduce δ in Proposition 4, leading to less catastrophic Goodhart.

5 Numerical Results

Prompts Dataset, LLMs and Reward Models We consider the attaq dataset (Kour et al., 2023) consisting of 1.4k prompts that triggers undesirable behaviors in LLMs. We use as Reward model FsfairX-LLaMA3-RM-v0.1 Dong et al. (2023); Xiong et al. (2024) this is among the best reward model for measuring helpfulness, safety, instruction following and lack of toxicity Lambert et al. (2024a). We use three LLMs from each we sample with top-p sampling and a temperature τ 100 responses for each prompt, models are: Merlinite Sudalairaj et al. (2024) (a base model not aligned), mixtral-8x7b-instruct Jiang et al. (2024) and llama-2-13b-chat Touvron et al. (2023) (aligned models with different reward).

Results and Discussion We plot in Figure 2 the histograms of the reward under the original LLM policy (first panel) and under the best of n policy for $n=10,50,100$ (second to fourth panels). We observe that the reference rewards are not heavy-tailed and follow sub-Gaussian/sub-Gamma distributions (appendix G provides tests such as q-q plots to probe this). Table 2 presents statistics of these distributions. We see that

Mixtral achieves the highest mean reward using the best of n for $n = 100$, while Merlinite shows the greatest improvement over the original policy

This is due to the fact that Merlinite is a base model and not an aligned one. In Figure 3, we plot the centered best of n rewards against the KL divergence of the best of n policy to the reference policy, and observe that it follows the form $a\sqrt{x} + bx$, indicating a sub-Gamma tail as shown in Proposition 3. Looking at the estimated coefficient a in Table 3 and the standard deviation of the reward under reference models in Table 2 (denoted as all), we observe that a is slightly larger than the standard deviation of the reward under the reference policy, which aligns with our theory. If we apply Pinsker's inequality, the upper bound on the reward would scale with the maximum reward, as also given in Table 2 , which is much larger than a, resulting in a loose bound. Thus, the reward/KL plots are governed by the tail properties of the reference policy.

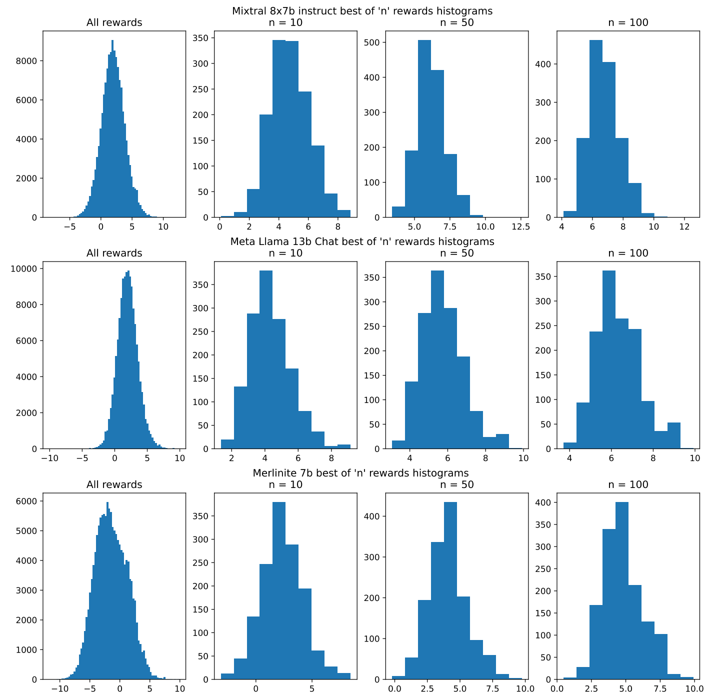

Figure 2: Histogram of reward and best of n reward with FsfairX-LLaMA3-RM-v0.1

Model and best of n	**Mean**	**Std**	**Max**
Merlinite all	-1.42	2.66	9.91
Merlinite n = 10	2.33	1.81	8.44
Merlinite n = 50	4.07	1.45	9.74
Merlinite n = 100	4.76	1.40	9.91
Mixtral all	1.96	1.86	12.57
Mixtral n = 10	4.74	1.31	8.86
Mixtral n = 50	6.21	1.03	12.57
Mixtral n = 100	6.78	0.98	12.57
Llama13bchat all	1.98	1.58	9.92
Llama13bchat n = 10	4.35	1.23	9.12
Llama13bchat n = 50	5.74	1.09	9.92
Llama13bchat n = 100	6.31	1.04	9.92

Table 2: Comparison of Mean, Std, and Max values for the reward model FsfairX-LLaMA3-RM-v0.1 evaluating samples from different models using various best of n policies.

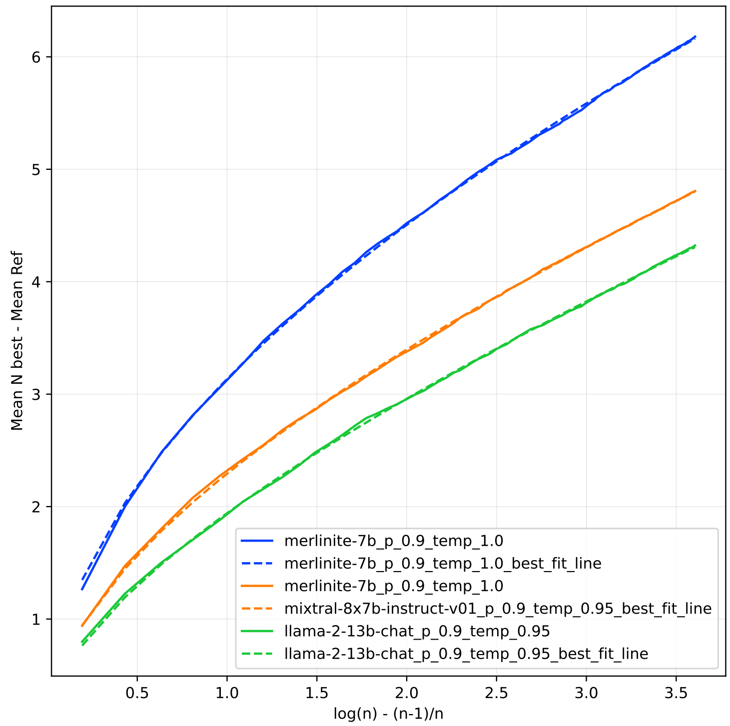

Figure 3: (y-axis) The centered reward using FsfairX-LLaMA3-RM-v0.1 reward model of best of n policy $\mathbb{E}_{\pi^n_{r,\text{ref}}} r - \mathbb{E}_{\pi_{\text{ref}}} r$ versus (x-axis) $\mathsf{KL}(\pi^n_{r,\text{ref}}, \pi_{\text{ref}})$ policy computed as $\log(n) - \frac{n-1}{n}$. We also plot fitted curves $y = a\sqrt{x} + bx$. Coefficients are given in Tab. 3.

Model	a	b
Merlinite-7B	3.01309273	0.12227016
Mixtral-8x7B-instruct	2.03057845	0.26265606
Llama 13B-chat	1.57822599	0.36367755

Table 3: Best fitted curves coefficients $y = a\sqrt{x} + bx$

Reward Versus Rényi divergence for Best of n Varying n, we compute the centered expected reward of the best of n policy versus the α Rényi divergence. The Rényi divergence of the best of n policy is computed using the expression given in Table 1. For $\alpha \in (0, 1)$ for smaller values of the Rényi divergence we achieve higher reward than KL (See Figures 16, 17 and 18). For $\alpha > 1$, for same value of KL, we achieve higher reward using KL than using the Rényi divergence for $\alpha > 1$, (See Figures 19, 20 and 21). This suggests that for same value of the divergence, α-Rényi divergence for $\alpha > 1$ may allow for less reward hacking. This observation was used in Huang et al. (2024) using the Chi-Squared regularizer ($\alpha = 2$) in addition to KL to fight reward overoptimization.

6 Conclusion

We presented in this paper a comprehensive information theoretical analysis of policy alignment using reward optimization with RL and best of n sampling. We showed for best of n a bound on KL under realistic assumptions on the reward. Our analysis showed that the alignment reward improvement, is intrinsically constrained by the tails of the reward under the reference policy and controlling the KL divergence results in an upper bound of the policy improvement. We showed that the KL bound may not be tight if the tails of the optimized policy satisfy a condition expressed via Rényi divergence. We also explained the deterioration of the golden reward via overestimation of the proxy reward.

References

Venkat Anantharam. A variational characterization of rényi divergences. *IEEE Transactions on Information Theory*, 64(11):6979–6989, 2018.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy, 2024.

Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Luc Rey-Bellet, and Jie Wang. Variational representations and neural network estimation of rényi divergences. *SIAM Journal on Mathematics of Data Science*, 3(4):1093–1116, 2021.

Stéphane Boucheron and Maud Thomas. Concentration inequalities for order statistics. 2012.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL `https://doi.org/10.1093/acprof:oso/9780199535255.001.0001`.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf`.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=dcjtMYkpXx`.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, and Marc Dymetman. Compositional preference models for aligning LMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=tiiAzqi6Ol`.

Jacob Hilton and Leo Gao. Measuring goodhart's law, 2022.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Gautam Kamath. Bounds on the expectation of the maximum of samples from a gaussian. *URL http://www. gautamkamath. com/writings/gaussian max. pdf*, 10:20–30, 2015.

George Kour, Marcel Zalmanovici, Naama Zwerdling, Esther Goldbraich, Ora Nova Fandina, Ateret Anaby-Tavor, Orna Raz, and Eitan Farchi. Unveiling safety vulnerabilities of large language models. *arXiv preprint arXiv:2311.04124*, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling. `https://huggingface.co/spaces/allenai/reward-bench`, 2024a.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024b.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Y. Polyanskiy and Y. Wu. Information theory: From coding to learning, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Alfréd Rényi. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4: 191–231, 1953. URL `https://api.semanticscholar.org/CorpusID:123132570`.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser, Cham, 2015. ISBN 9783319208275. doi: 10.1007/978-3-319-20828-2.

Ofer Shayevitz. On rényi measures and hypothesis testing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pp. 894–898, 2011. doi: 10.1109/ISIT.2011.6034266.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D Cox, and Akash Srivastava. Lab: Large-scale alignment for chatbots. *arXiv preprint arXiv:2403.01081*, 2024.

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

Tim van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024a. URL `https://openreview.net/forum?id=2cRzmWXK9N`.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alex D'Amour, Sanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models. *arXiv preprint arXiv:2402.00742*, 2024b.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint, 2024.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment, 2024.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.276. URL `https://aclanthology.org/2021.naacl-main.276`.

Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. Calibrating sequence likelihood improves conditional language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=0qSOodKmJaN`.

A Broader Impact and Limitations

We believe this work explaining scaling laws for reward models and alignment will give practitioners insights regarding the limits of what is attainable via alignment. All assumptions under which our statements hold are given. We don't see any negative societal impact of our work.

B Proofs For Best of n Policy

B.1 Best of n Policy KL Guarantees

Proof 1 (Proof of Lemma 1)

$$\mathsf{KL}(E^{(n)}||E) = \int_0^{+\infty} f_{E^{(n)}}(x) \log\left(\frac{f_{E^{(n)}}(x)}{f_E(x)}\right) dx$$

We have $f_E(x) = e^{-x}1_{x\geq 0}$. *Note that the CDF of maximum of exponential* $F_{E^{(n)}}(x) = (1-e^{-x})1_{x\geq 0}$, *and hence* $f_{E^{(n)}}(x) = n(1-e^{-x})^{n-1}e^{-x}1_{x\geq 0}$. *Hence we have:*

$$\begin{aligned}\mathsf{KL}(E^{(n)}||E) &= \int_0^{+\infty} n(1-e^{-x})^{n-1}e^{-x}\log\left(\frac{n(1-e^{-x})^{n-1}e^{-x}}{e^{-x}}\right)dx\\ &= \int_0^{+\infty} n(1-e^{-x})^{n-1}e^{-x}\log\left(n(1-e^{-x})^{n-1}\right)dx\end{aligned}$$

Let $u = 1-e^{-x}$, *we have* $du = e^{-x}dx$. *It follows that :*

$$\begin{aligned}\mathsf{KL}(E^{(n)}||E) &= \int_0^1 nu^{n-1}\log\left(nu^{n-1}\right)du\\ &= \int_0^1 nu^{n-1}\left(\log(n)+(n-1)\log(u)\right)du\\ &= \log(n)\int_0^1 du^n + (n-1)\int_0^1 nu^{n-1}\log(u)du\\ &= \log(n) + (n-1)\int_0^1 d(u^n\log u - \frac{u^n}{n})\\ &= \log(n) - \frac{n-1}{n}.\end{aligned}$$

Proof 2 (Proof of Theorem 1) *Recall that* $T_X = F^{-1}_{R(Y)|X}\circ F_E$, F_E *is one to one. If the space* $\mathcal{Y}$ *is finite,* $R(Y|X)$ *has a discontinuous CDF hence not strictly monotonic. It follows that its quantile* $F^{-1}_{R(Y)|X}$ *is not a one to one map and* T_X *as a result is not a one to one map and hence we have by DPI (that is an inequality in this case since* T_X *is not one to one):*

$$\mathsf{KL}(T_X(E^{(n)})||T_X(E)) \leq \mathsf{KL}(E^{(n)}||E) \tag{18}$$

If the space $\mathcal{Y}$ *is infinite and we assume that* $R(Y|X)$ *is continuous and strictly monotonic then* $F^{-1}_{R(Y)|X}$ *is a one to one map, and as a result* T_X *is a one to one map and the DPI is an equality in this case:*

$$\mathsf{KL}(T_X(E^{(n)})||T_X(E)) = \mathsf{KL}(E^{(n)}||E) \tag{19}$$

Hence under Assumption 1 and for $\mathcal{Y}$ *finite combining* (11) *and* (18) *we have:*

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{ref}|X) \leq \mathsf{KL}(E^{(n)}||E), \tag{20}$$

and under Assumption 1 and for $\mathcal{Y}$ *infinite and assuming* $F_{R(Y)|X}$ *is continuous and strictly monotonic, combining* (11) *and* (19) *we have:*

$$\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{ref}|X) = \mathsf{KL}(E^{(n)}||E). \tag{21}$$

Under the more realistic Assumption 2 we can also apply the DPI on the stochastic map H_X, since DPI also holds for stochastic maps (under our assumption $R|X \to Y|X$ see for example (van Erven & Harremos, 2014) Example 2)

$$\begin{aligned}&\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{ref}|X) = \mathsf{KL}(H_X(R^n(Y))||H_X(R(Y))|X))\\&\leq \mathsf{KL}(R^n(Y)||R(Y)|X) = \mathsf{KL}(T_X(E^{(n)})||T_X(E)),\end{aligned} \tag{22}$$

and hence under Assumption 2 regardless whether T_X is a one to one map or not, thus we have: $\mathsf{KL}(\pi^{(n)}_{r,\text{ref}}||\pi_{ref}|X) \leq \mathsf{KL}(E^{(n)}||E)$.

B.2 Best of n Policy f divergence and Rényi Divergence

Best of n Policy f divergence and Renyi divergence Guarantees Given that our proof technique relies on DPI and Rényi representation, we show that similar results hold for any f-divergence and for the Rényi divergence:

$$D_f\left(P||Q\right) = \int q(x) f\left(\frac{p(x)}{q(x)}\right)dx, \tag{23}$$

where f is convex and $f(1) = 0$. Hence we have by DPI for f-divergences:

Theorem 5 *Under Assumption 2 the best of n policy satisfies for any f-divergence:*

$$\mathrm{D}_f(\pi^{(n)}_{r,\text{ref}}||\pi_{\text{ref}}) \leq \int_0^1 f\left(nu^{n-1}\right)du \tag{24}$$

Proof 3 (Proof of Theorem 5)

$$\begin{aligned}\mathrm{D}_f(\pi^{(n)}_{r,\text{ref}}||\pi_{ref}|X) &= \mathrm{D}_f(Y^{(n)}||Y|X)\\&= \mathrm{D}_f(H_X(R^n(Y))||H_X(R(Y))|X)\\&\leq \mathrm{D}_f(R^n(Y)||R(Y)|X) \textit{ By the data processing inequality} &(25)\\&= \mathrm{D}_f(T_X(E^{(n)})||T_X(E)) \textit{ Renyi and Optimal Transport Representations } (8)\\&= \mathrm{D}_f(E^{(n)}||E) \textit{ since } T_X \textit{ is a monotonic bijection DPI is an equality} &(26)\\&= \int_0^{+\infty} f_E(x) f\left(\frac{f_{E^{(n)}}(x)}{f_E(x)}\right)dx &(27)\\&= \int_0^{\infty}(e^{-x}) f\left(n(1-e^{-x})^{n-1}\right)du &(28)\\&= \int_0^1 f(nu^{n-1})du. &(29)\end{aligned}$$

In particular we have the following bounds for common f divergences:

- *For $f(x) = x\log(x)$ we obtain the KL divergence and we have the result:*

$$\int_0^1 nu^{n-1}\log(nu^{n-1})du = \mathsf{KL}(E^{(n)}||E) = \log(n) - \frac{n-1}{n}.$$

- *For $f(x) = (x-1)^2$ we obtain the chi-squared divergence and we have:* $\int_0^1 \left(nu^{n-1}-1\right)^2 du = \int_0^1 (n^2u^{2(n-1)} - 2nu^{n-1} + 1)du = \frac{n^2}{2n-1}u^{2n-1} - 2u^n + u|_0^1 = \frac{n^2}{2n-1} - 2 + 1 = \frac{n^2-2n+1}{2n-1} = \frac{(n-1)^2}{2n-1}$.

- *For* $f(x) = \frac{1}{2}|x-1|$, *we obtain the total variation distance (TV) and we have:* $\frac{1}{2}\int_0^1 \left|nu^{n-1}-1\right| du = \frac{1}{2}(\int_0^{u^*}\left(1-nu^{n-1}\right)du + (\int_{u^*}^1 \left(nu^{n-1}-1\right)du) = (u^* - (u^*)^n)$,*where* $n(u^*)^{(n-1)} = 1$, *i.e* $u^* = (\frac{1}{n})^{\frac{1}{n-1}}$ *. Hence the TV is* $(\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}$.

- *For* $f(x) = (1-\sqrt{x})^2$ *we have the hellinger distance:* $\int_0^1\left(\sqrt{n}u^{\frac{n-1}{2}}-1\right)^2 du = \int_0^1 (nu^{n-1} - 2\sqrt{n}u^{\frac{n-1}{2}} + 1)du = u^n - 2\sqrt{n}\frac{u^{\frac{n+1}{2}}}{\frac{n+1}{2}} + u\Big|_0^1 = 2(1-\frac{2\sqrt{n}}{n+1}) = 2\frac{(1-\sqrt{n})^2}{n+1}$

- *For* $f(x) = -\log(x)$, *we obtain the forward KL and we have :* $\int_0^1 f(nu^{n-1})du = n-1-\log(n)$.

Guarantees with Rényi Divergence Turning now to the Rényi divergence for $\alpha \in (0,1)\cup(1,\infty)$:

$$D_\alpha(P||Q) = \frac{1}{(\alpha-1)}\log\left(\int p^\alpha(x)q^{1-\alpha}(x)dx\right)$$

the limit as $\alpha \to 1$ $D_1(P||Q)) = \mathsf{KL}(P||Q)$.

Theorem 6 *Under Assumption 2 the best of n policy satisfies:*

$$D_\alpha(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) \leq \frac{1}{(\alpha-1)}\log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right) \tag{30}$$

Proof 4 (Proof of Theorem 6) *Applying DPI that holds also for the Rényi divergence twice from* $Y, Y^{(n)}$ *to* $R, R^{(n)}$ *and from* $R, R^{(n)}$ *to* $E, E^{(n)}$ *we obtain :*

$$D_\alpha(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{ref}|X) \leq D_\alpha(E^{(n)}||E)$$

$$\begin{aligned} D_\alpha(E^{(n)}||E) &= \frac{1}{(\alpha-1)}\log\left(\int_0^\infty n^\alpha(1-e^{-x})^{\alpha(n-1)}e^{-\alpha x}e^{-x(1-\alpha)}dx\right) \\ &= \frac{1}{(\alpha-1)}\log\left(\int_0^{+\infty} n^\alpha(1-e^{-x})^{\alpha(n-1)}e^{-x}dx\right) \end{aligned}$$

Let $u = 1-e^{-x}$ *we have* $du = e^{-x}dx$

$$\begin{aligned} D_\alpha(E^{(n)}||E) &= \frac{1}{(\alpha-1)}\log\left(\int_0^1 n^\alpha u^{\alpha(n-1)}du\right) \\ &= \frac{1}{(\alpha-1)}\left(\log n^\alpha + \log\int_0^1 u^{\alpha(n-1)}du\right) \\ &= \frac{1}{(\alpha-1)}\left(\log n^\alpha + \log\frac{u^{\alpha(n-1)+1}}{\alpha(n-1)+1}\Big|_0^1\right) \\ &= \frac{1}{(\alpha-1)}\log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right) \end{aligned}$$

From Renyi to KL guarantees Let $s_1(\alpha) = (\alpha-1)$, and $s_2(\alpha) = \log\left(\frac{n^\alpha}{\alpha(n-1)+1}\right)$, we have $D_\alpha(E^{(n)}||E) = \frac{s_2(\alpha)}{s_1(\alpha)}$, we have $\mathsf{KL}(E^{(n)}||E) = \lim_{\alpha\to 1} D_\alpha(E^{(n)}||E) = \lim_{\alpha\to 1}\frac{s_2(\alpha)}{s_\alpha} = \frac{0}{0}$, hence applying L'Hôpital rule we have: $\lim_{\alpha\to 1}\frac{s_2(\alpha)}{s_1(\alpha)} = \lim_{\alpha\to 1}\frac{s_2'(\alpha)}{s_1'(\alpha)} = \lim_{\alpha\to 1}\frac{\log(n)-\frac{n-1}{\alpha(n-1)+1}}{1} = \log(n) - \frac{n-1}{n}$. Hence we recover the result for the KL divergence.

B.3 Best of n Dominance

Proof 5 (Proof of Proposition 1) *$F_{E^{(n)}}(x) = (F_E(x))^n \leq F_E(x), \forall x \geq 0$, which means also that $F^{-1}_{E^{(n)}}(t) \geq F_E^{-1}(t), \forall t \in [0,1]$, which means that $E^{(n)}$ dominates E in the first stochastic order : $E^{(n)} \underset{FSD}{\succcurlyeq} E$, which means there exists a coupling between $E^{(n)}$ and E, $\pi \in \Pi(E^{(n)}, E)$, such that $E \geq e$, for all $(E,e) \sim \pi$. On the other hand By Rényi and Monge map representations we have: $R^{(n)} = F_R^{-1} \circ F_E(E^{(n)})$ and $R = F_R^{-1} \circ F_E(E)$, given that $T = F_R^{-1} \circ F_E$ is non decreasing the same coupling π guarantees that $T(E) \geq T(e)$, for all $(E,e) \sim \pi$ and Hence $R^{(n)} \underset{FSD}{\succcurlyeq} R$.*

Corollary 2 *Best of n-polciy has higher expectation :*

$$\mathbb{E}R^{(n)} \geq \mathbb{E}R,$$

and is a safer policy, let the Tail Value at Risk be:

$$\mathrm{TVAR}_p(X) = \frac{1}{p}\int_0^p Q_R(t)dt$$

We have

$$\mathrm{TVAR}_p(R^n) \geq \mathrm{TVAR}_p(R), \forall p \in [0,1]$$

Proof 6 (Proof of Corollary 2) *First order dominance implies second order dominance (i.e by integrating quantiles). Expectation is obtained for $p = 1$.*

C Best of n and RL Policy

Proof 7 (Proof of Proposition 2) *We fix here $\beta = \frac{1}{\lambda_\Delta}$*

$$\begin{aligned}\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\beta,r}) &= \int \pi^{(n)}_{r,\mathrm{ref}}(y|x)\log\left(\frac{\pi^{(n)}_{r,\mathrm{ref}}(y|x)}{\pi_{\beta,r}(y|x)}\right) = \int \pi^{(n)}_{r,\mathrm{ref}}(y|x)\log\left(\frac{\pi^{(n)}_{r,\mathrm{ref}}(y|x)}{\pi_{\mathrm{ref}}(y|x)\frac{e^{\beta r(x,y)}}{Z_\beta(x)}}\right)\\ &= \mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) + \log\left(\mathbb{E}_{\pi_{\mathrm{ref}}}e^{\beta r}\right) - \beta\int r d\pi^{(n)}_{r,\mathrm{ref}}\end{aligned}$$

On the other hand by optimality of $\pi_{\beta,r}$ we have:

$$\mathsf{KL}\left(\pi_{\beta,r}||\pi_{\mathrm{ref}}\right) = \beta\int r d\pi_{\beta,r} - \log\left(\int e^{\beta r}d\pi_{\mathrm{ref}}\right)$$

and hence we have:

$$\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\beta,r}) = \mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) - \mathsf{KL}\left(\pi_{\beta,r}||\pi_{\mathrm{ref}}\right) + \beta\left(\int r d\pi_{\beta,r} - \int r d\pi^{(n)}_{r,\mathrm{ref}}\right)$$

We choose n such that :

$$\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\mathrm{ref}}) \leq \log(n) - \frac{n-1}{n} \leq \mathsf{KL}\left(\pi_{\beta,r}||\pi_{\mathrm{ref}}\right) = \Delta$$

and we conclude choosing $n = e^\Delta$ therefore for that choice of n that:

$$\mathsf{KL}(\pi^{(n)}_{r,\mathrm{ref}}||\pi_{\beta,r}) \leq \beta\left(\int r d\pi_{\beta,r} - \int r d\pi^{(n)}_{r,\mathrm{ref}}\right)$$

On the other hand we have:

$$\begin{aligned}
\left|\int r d\pi_{\beta,r} - \int r d\pi^{(n)}_{r,\text{ref}}\right| &= \left|\int r\exp(\beta r)\frac{1}{Z_\beta}d\pi_{\text{ref}} - \int \max_i r(x_i)d\pi_{\text{ref}}(x_1)\dots d\pi_{\text{ref}}(x_n)\right| \\
&= \left|\int\left(\frac{1}{n}\sum_{i=1}^{n}\frac{r(x_i)\exp(\beta r(x_i))}{Z_\beta} - \max_i r(x_i)\right)d\pi_{\text{ref}}(x_1)\dots d\pi_{\text{ref}}(x_n)\right| \\
&= \left|\int\left(\frac{1}{n}\sum_{i=1}^{n}\frac{r(x_i)\exp(\beta r(x_i))}{\sum_{i=1}^{n}\exp(\beta r(x_i))}\frac{\sum_{i=1}^{n}\exp(\beta r(x_i))}{Z_\beta} - \max_i r(x_i)\right)d\pi_{\text{ref}}(x_1)\dots d\pi_{\text{ref}}(x_n)\right| \\
&\leq \int\left|\max r(x_i)\left(\frac{\frac{1}{n}\sum_{i=1}^{n}\exp(\beta r(x_i))}{Z_\beta} - 1\right)\right|d\pi_{\text{ref}}(x_1)\dots d\pi_{\text{ref}}(x_n) \\
&\leq \frac{M}{Z_\beta}\mathbb{E}\left|\sum_{i=1}^{n}\exp(\beta r(x_i)) - Z_\beta\right|
\end{aligned}$$

where we used the following fact, followed by Jensen inequality :

$$\sum_{i=1}^{n}\frac{r(x_i)\exp(\beta r(x_i))}{\sum_{i=1}^{n}\exp(\beta r(x_i))} \leq \max_i r(x_i).$$

Assume that the reward is bounded hence we have by Hoeffding inequality :

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\exp(\beta r(x_i)) - Z_\beta\right| \geq t\right) \leq 2e^{-\frac{nt^2}{2(\exp(\beta M)-\exp(-\beta M))^2}}$$

Hence we have:

$$\mathbb{E}\left|\sum_{i=1}^{n}\exp(\beta r(x_i)) - Z_\beta\right| \leq 2\sqrt{\frac{\pi}{2}}\frac{\exp(\beta M)-\exp(-\beta M)}{\sqrt{n}}$$

$$\mathsf{KL}(\pi^{(\exp(\Delta))}_{r,\text{ref}}||\pi_{\lambda_\Delta,r}) \leq \frac{M}{\lambda_\Delta Z_{1/\lambda_\Delta}}\sqrt{2\pi}(\exp(\beta M)-\exp(-\beta M))\sqrt{\exp(-\Delta)}.$$

D Transportation Inequalities and KL Divergence

D.1 Transportation Inequalities with KL

The following Lemma (Lemma 4.14 in (Boucheron et al., 2013)) uses the Donsker-Varadhan representation of the KL divergence to obtain bounds on the change of measure , and using the tails of π_{ref}.

Lemma 4 (Lemma 4.14 in (Boucheron et al., 2013)) *Let ψ be a convex and continuously differentiable function ψ on a possibly unbounded interval $[0, b)$, and assume $\psi(0) = \psi'(0) = 0$. Define for every $x \geq 0$, the convex conjugate $\psi^*(x) = \sup_{\lambda\in[0,b)} \lambda x - \psi(\lambda)$, and let $\psi^{*-1}(t) = \inf\{x \geq 0 : \psi^*(x) > t\}$. Then the following statements are equivalent:*
(i) For $\lambda \in [0, b)$

$$\log\left(\int e^{\lambda(r-\int r dQ)}dQ\right) \leq \psi(\lambda),$$

(ii) For any probability measure P that is absolutely continuous with respect to Q and such that $\mathsf{KL}(P||Q) < \infty$:

$$\int r dP - \int r dQ \leq \psi^{*-1}(\mathsf{KL}(P||Q)).$$

Lemma 5 (Inverse of the conjugate (Boucheron et al., 2013)) *1. If $Q \in \mathsf{SubGauss}(\sigma^2)$, we have for $t \geq 0$ $\psi^{*-1}(t) = \sqrt{2\sigma^2 t}$.*

2. *If $Q \in \mathsf{Subgamma}(\sigma^2, c)$, we have for $t \geq 0$ $\psi^{*-1}(t) = \sqrt{2\sigma^2 t} + ct$.*

We give here a direct proof for the subgaussian case:

Proof 8 *By the Donsker Varadhan representation of the* KL *we have:*

$$\mathsf{KL}(P||Q) = \sup_h \int h dP - \log\left(\int e^h dQ\right)$$

Fix x and $M > 0$ and define for $0 < \lambda < M$

$$h_\lambda(y) = \lambda\left(r(x,y) - \mathbb{E}_{\pi_{\text{ref}}(y|x)} r(x,y)\right)$$

We omit in what follows x and y, but the reader can assume from here on that π and π_{ref} are conditioned on x. Note that $R_{\text{ref}}|x = (r(x,.))_\sharp \pi_{\text{ref}}(.|x)$ and we assume $R_{\text{ref}}|x$ subgaussian. Note that

$$\mathbb{E}_{\pi_{\text{ref}}} e^{h_\lambda} = \mathbb{E}_{\pi_{\text{ref}}|x} e^{\lambda(r - \mathbb{E}_{\pi_{\text{ref}}|x} r)} = M_{R_{\text{ref}}|x}(\lambda),$$

where $M_{R_{\text{ref}}|x}$ the moment generating function of the reward under the reference policy. $R_{\text{ref}}|x$ is subgaussian we have for all $\lambda \in \mathbb{R}$:

$$\mathbb{E}_{\pi_{\text{ref}}|x} e^{h_\lambda} \leq e^{\frac{\lambda^2\sigma^2}{2}} \leq e^{\frac{M^2\sigma^2}{2}} < \infty$$

Hence $h_\lambda \in \mathcal{H}$ and we have for all $\pi \lll \pi_{\text{ref}}$ and for all $0 < M < \infty$ and $0 < \lambda < M$:

$$\lambda \mathbb{E}_{\pi|x}(r - \mathbb{E}_{\pi_{\text{ref}}|x} r) \leq \mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \log\left(\mathbb{E}_{\pi_{\text{ref}}|x} e^{\lambda(r - \mathbb{E}_{\pi_{\text{ref}}|x} r)}\right)$$

or equivalently:

$$\mathbb{E}_{\pi|x} r - \mathbb{E}_{\pi_{\text{ref}}|x} r \leq \frac{1}{\lambda}\mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \frac{1}{\lambda}\log\left(\mathbb{E}_{\pi_{\text{ref}}|x} e^{\lambda(r - \mathbb{E}_{\pi_{\text{ref}}|x} r)}\right)$$

Finally we have for $\pi \lll \pi_{\text{ref}}$ for all $0 < \lambda < M$:

$$\mathbb{E}_{\pi|x} r - \mathbb{E}_{\pi_{\text{ref}}|x} r \leq \frac{1}{\lambda}\mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \frac{1}{\lambda}\log\left(M_{R_{\text{ref}}|x}(\lambda)\right) \tag{31}$$

Being a subgaussian, the MGF of $R_{\text{ref}}|x$ is bounded as follows:

$$\log\left(M_{R_{\text{ref}}|x}(\lambda)\right) \leq \frac{\lambda^2\sigma^2}{2}.$$

Hence we have for :

$$\mathbb{E}_{\pi|x} r - \mathbb{E}_{\pi_{\text{ref}}|x} r \leq \frac{1}{\lambda}\mathsf{KL}(\pi||\pi_{\text{ref}}|x) + \frac{\lambda\sigma^2}{2}$$

Integrating over x we obtain for all $\pi \lll \pi_{\text{ref}}$ and all $0 < \lambda < M$:

$$\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \frac{1}{\lambda}\mathsf{KL}(\pi||\pi_{\text{ref}}) + \frac{\lambda\sigma^2}{2}$$

Define :

$$\delta(\lambda) = \frac{1}{\lambda}\mathsf{KL}(\pi||\pi_{\text{ref}}) + \frac{\lambda\sigma^2}{2}$$

minimizing the upper bound $\delta(\lambda)$ for $\lambda \in (0, M]$, taking derivative $\delta'(\lambda) = -\frac{\mathsf{KL}(\pi||\pi_{\text{ref}})}{\lambda^2} + \frac{\sigma^2}{2} = 0$ gives $\lambda^* = \sqrt{\frac{2\mathsf{KL}(\pi||\pi_{\text{ref}})}{\sigma^2}}$. Taking $M = 2\lambda^*$, λ^* is the minimizer. Putting this in the bound we have finally for all rewards r for all π:

$$\mathbb{E}_\pi r - \mathbb{E}_{\pi_{\text{ref}}} r \leq \sqrt{2\sigma^2 \mathsf{KL}(\pi||\pi_{\text{ref}})}. \tag{32}$$

Proof 9 (Proof of Corollary 1) *(i) This follows from optimality of π_{λ_Δ} and applying the transportation inequality for gaussian tail.*

(ii) This follows from applying Corollary 2 (best of n policy has larger mean) and Theorem 1 for bounding the KL.

Proof 10 (Proof of Theorem 2) *For the penalized RL we have by optimality:*

$$\begin{aligned}\int r d\pi_{\beta,r} - \frac{1}{\beta}\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}}) &= \frac{1}{\beta}\log\left(\int e^{\beta r} d\pi_{\text{ref}}\right)\\ &= \frac{1}{\beta}\log\left(\int e^{\beta(r-\int r d\pi_{\text{ref}})} d\pi_{\text{ref}}\right) + \int r d\pi_{\text{ref}}\end{aligned}$$

It follows that :

$$\frac{1}{\beta}\log\left(\int e^{\beta(r-\int r d\pi_{\text{ref}})} d\pi_{\text{ref}}\right) = \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} - \frac{1}{\beta}\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}}) \tag{33}$$

On the other hand by the variational representation of the Rényi divergence we have:

$$\begin{aligned}\int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} \leq \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta} - \frac{1}{\beta-1}\log\left(\int e^{(\beta-1)(r-\int r d\pi_{\beta,r})} d\pi_{\beta,r}\right)\\ + \frac{1}{\beta}\log\left(\int e^{\beta(r-\int r d\pi_{\text{ref}})} d\pi_{\text{ref}}\right)\end{aligned} \tag{34}$$

Summing Equations (33) *and* (34) *we obtain a bound on the moment generating function at β of $r_\sharp\pi_{\beta,r}$ (this is not a uniform bound , it holds only for β):*

$$\frac{1}{\beta-1}\log\left(\int e^{(\beta-1)(r-\int r d\pi_{\beta,r})} d\pi_{\beta,r}\right) \leq \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta}. \tag{35}$$

Let us assume $\beta > 1$ we have therefore the following bound on the logarithmic moment generation function at $\beta - 1$

$$\psi_{r_\sharp\pi_{\beta,r}}(\beta-1) \leq \frac{\beta-1}{\beta}\left(D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})\right)$$

Let $R_{i,\beta} = r_\sharp\pi_{\beta,r}, i = 1\ldots m$, the reward evaluation of m independent samples of $\pi_{\beta,r}$ we have:

$$\begin{aligned}\mathbb{P}\Big\{\sum_{i=1}^m (R_{i,\beta} - \int r d\pi_{\beta,r}) > mt\Big\} &= \mathbb{P}(e^{\sum_{i=1}^m (\beta-1)(R_{i,\beta}-\int r d\pi_{\beta,r})} > e^{m(\beta-1)t})\\ &\leq e^{-(\beta-1)mt} e^{m\psi_{R_\beta}(\beta-1)}\\ &\leq e^{-(\beta-1)mt} e^{m\frac{\beta-1}{\beta}(D_\beta(\pi_{\beta,r}||\pi_{\text{ref}})-\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}}))}\\ &\leq e^{-m(\beta-1)\left(t-\frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}})-\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta}\right)}\end{aligned} \tag{36}$$

Let $t_0 > 0$, hence we have for $\beta > 1$:

$$\mathbb{P}\Big\{\frac{1}{m}\sum_{i=1}^m R_{i,\beta} > \int r d\pi_{\beta,r} + t_0 + \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta}\Big\} \leq e^{-m(\beta-1)t_0}$$

Now turning to $R_{\text{ref}} = r_\sharp\pi_{\text{ref}}$, since $R_{\text{ref}} \in \mathsf{SubGauss}(\sigma^2_{\text{ref}})$ we have for every $t_0 > 0$:

$$\mathbb{P}\Big\{-\frac{1}{m}\sum_{i=1}^m R_{i,\text{ref}} > -\int r d\pi_{\text{ref}} + t_0\Big\} \leq e^{-\frac{mt_0^2}{2\sigma^2_{\text{ref}}}}$$

Hence we have with probability at least $1 - e^{-\frac{mt_0^2}{2\sigma_{\text{ref}}^2}} - e^{-m(\beta-1)t_0}$:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} R_{i,\beta} - \frac{1}{m}\sum_{i=1}^{m} R_{i,\text{ref}} &\leq \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} + 2t_0 + \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta} \\ &\leq \sqrt{2\sigma_{\text{ref}}^2 \mathsf{KL}(\pi||\pi_{\text{ref}})} + 2t_0 + \frac{D_\beta(\pi_{\beta,r}||\pi_{\text{ref}}) - \mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})}{\beta}.\end{aligned}$$

E Proofs for Transportation Inequalities and Rényi Divergence

Proposition 6 (Fenchel Conjugate Propreties) *Let F and G be convex functions on a space E and F^*, G^* be their convex conjugates defined on E^*. We have:*

1. *Let $F_\gamma(x) = \gamma F(x)$ we have:*
$$F_\gamma^*(p) = \gamma F^*\left(\frac{p}{\gamma}\right) \tag{37}$$
2. *Duality:*
$$\min_{x\in E} F(x) + G(x) = \max_{p\in E^*} -F^*(-p) - G^*(p) \tag{38}$$
3. *Toland Duality:*
$$\min_{x\in E} F(x) - G(x) = \min_{p} G^*(p) - F^*(p) \tag{39}$$

Proof 11 (Proof of Theorem 3) *Let $\gamma > 0$, let $F_{P,\gamma}(R) = \gamma \mathsf{KL}(R||P)$, the Fenchel conjugate of $F_{P,1}(.)$ is defined for h bounded and measurable function as follows $F_{P,1}^*(h) = \log \mathbb{E}_P e^h$. It follows by 1) in Proposition 6 that : $F_{P,\gamma}^*(h) = \gamma F_{P,1}^*(\frac{h}{\gamma}) = \gamma \log \mathbb{E}_P e^{\frac{h}{\gamma}}$.*
For $0 < \alpha < 1$: The objective function in (15) *is the sum of convex functions: $F_{P,\alpha}(R) + F_{Q,1-\alpha}(R)$, by (2) in Proposition 6, we have by duality:*

$$\begin{aligned}(1-\alpha)D_\alpha(P||Q) &= \inf_R F_{P,\alpha}(R) + F_{Q,1-\alpha}(R) \\ &= \sup_{h\in\mathcal{H}} -F_{P,\alpha}^*(-h) - F_{Q,1-\alpha}^*(h) \\ &= \sup_{h\in\mathcal{H}} -\alpha \log \mathbb{E}_P e^{-\frac{h}{\alpha}} - (1-\alpha)\log \mathbb{E}_Q e^{\frac{h}{1-\alpha}}\end{aligned}$$

Replacing h by $(1-\alpha)(\alpha)h$ does not change the value of the sup and hence we obtain:

$$\begin{aligned}(1-\alpha)D_\alpha(P||Q) &= \sup_{h\in\mathcal{H}} -\alpha \log \mathbb{E}_P e^{-\frac{(1-\alpha)(\alpha)h}{\alpha}} - (1-\alpha)\log \mathbb{E}_Q e^{\frac{(1-\alpha)(\alpha)h}{1-\alpha}} \\ &= \sup_{h\in\mathcal{H}} -\alpha \log \mathbb{E}_P e^{-(1-\alpha)h} - (1-\alpha)\log \mathbb{E}_Q e^{\alpha h}.\end{aligned}$$

dividing by $\frac{1}{\alpha(1-\alpha)}$ both sides we obtain for $0 < \alpha < 1$:

$$\frac{1}{\alpha}D_\alpha(P||Q) = \sup_{h\in\mathcal{H}} -\frac{1}{1-\alpha}\log \mathbb{E}_P e^{-(1-\alpha)h} - \frac{1}{\alpha}\log \mathbb{E}_Q e^{\alpha h}$$

For $\alpha > 1$: The objective function in (15) *is the difference of convex functions: $F_{P,\alpha}(R) - F_{Q,\alpha-1}(R)$, by Toland Duality (3) in Proposition 6 we have:*

$$\begin{aligned}(1-\alpha)D_\alpha(P||Q) &= \inf_R F_{P,\alpha}(R) - F_{Q,\alpha-1}(R) \\ &= \inf_{h\in\mathcal{H}} F_{Q,\alpha-1}^*(h) - F_{P,\alpha}^*(h) \\ &= \inf_{h\in\mathcal{H}} (\alpha-1)\log \mathbb{E}_Q e^{\frac{h}{(\alpha-1)}} - \alpha \log \mathbb{E}_P e^{\frac{h}{\alpha}}\end{aligned}$$

The inf does not change when we replace h by $\alpha(\alpha-1)h$, hence we have:

$$\begin{aligned}(\alpha-1)D_\alpha(P||Q) &= -\inf_{h\in\mathcal{H}}(\alpha-1)\log\mathbb{E}_Q e^{\frac{\alpha(\alpha-1)h}{(\alpha-1)}} - \alpha\log\mathbb{E}_P e^{\frac{\alpha(\alpha-1)h}{\alpha}}\\ &= \sup_{h\in\mathcal{H}}\alpha\log\mathbb{E}_P e^{(\alpha-1)h} - (\alpha-1)\log\mathbb{E}_Q e^{\alpha h}\end{aligned}$$

dividing both sides by $\frac{1}{\alpha(\alpha-1)}$ we obtain for $\alpha > 1$:

$$\frac{1}{\alpha}D_\alpha(P||Q) = \sup_{h\in\mathcal{H}}\frac{1}{\alpha-1}\log\mathbb{E}_P e^{(\alpha-1)h} - \frac{1}{\alpha}\log\mathbb{E}_Q e^{\alpha h}.$$

Proof 12 (Proof of Lemma 2) *Adding and subtracting in the exponential $\int hdP$ and $\int hdQ$ resp we obtain the result:* $\frac{1}{\alpha-1}\log\left(\int e^{(\alpha-1)h}dP\right) - \frac{1}{\alpha}\log\left(\int e^{\alpha h}dQ\right) = \frac{1}{\alpha-1}\log\left(\int e^{(\alpha-1)(h-\int hdP+\int hdP)}dP\right) - \frac{1}{\alpha}\log\left(\int e^{\alpha(h-\int hdQ+\int hdQ)}dQ\right) = \int hdP - \int hdQ + \frac{1}{\alpha-1}\log\left(\int e^{(\alpha-1)(h-\int hdP)}dP\right) - \frac{1}{\alpha}\log\left(\int e^{\alpha(h-\int hdQ)}dQ\right)$

Proof 13 (Proof of Lemma 3) *Note that we have for $0<\alpha<1$, $\frac{1}{\alpha}D_\alpha(P||Q) = \frac{1}{1-\alpha}D_{1-\alpha}(Q||P)$ (See Proposition 2 in van Erven & Harremos (2014)). Taking limits we obtain* $\lim_{\alpha\to 0}\frac{1}{\alpha}D_\alpha(P||Q) = D_1(Q||P) = \mathsf{KL}(Q||P)$.

Proof 14 (Proof of Theorem 4) *For $0<\alpha<1$, we have for all $h\in\mathcal{H}$:*

$$\int hdP - \int hdQ \leq \frac{1}{\alpha}D_\alpha(P||Q) + \frac{1}{1-\alpha}\log\left(\int e^{(\alpha-1)(h-\int hdP)}dP\right) + \frac{1}{\alpha}\log\left(\int e^{\alpha(h-\int hdQ)}dQ\right) \tag{40}$$

Assuming r is bounded $0<r<b$ then we have $(r)_\sharp P - \mathbb{E}_P r$ and $(r)_\sharp Q - \mathbb{E}_Q r$ are sub-Gaussian with parameter $\sigma^2 = \frac{b^2}{4}$. Hence we have for $\lambda\in\mathbb{R}$:

$$\mathbb{E}_P e^{\lambda(r-\int rdP)} \leq \exp\left(\frac{\lambda^2\sigma_P^2}{2}\right) \text{ and } \mathbb{E}_Q e^{\lambda(r-\int rdQ)} \leq \exp\left(\frac{\lambda^2\sigma_Q^2}{2}\right),$$

Fix a finite $M>0$. For $0<\lambda<M$ and $P=\pi|x$ and $Q=\pi_{\text{ref}}|x$, consider $h_\lambda=\lambda r$, thanks to subgaussianity and boundedness of λ, $h_\lambda\in\mathcal{H}$ for all $\lambda\in(0,M)$. Hence we have by Equation (40) for all $\lambda\in(0,M)$:

$$\lambda\left(\int rdP - \int rdQ\right) \leq \frac{1}{\alpha}D_\alpha(P||Q) + \frac{1}{1-\alpha}\log\left(\int e^{\lambda(\alpha-1)(r-\int rdP)}dP\right) + \frac{1}{\alpha}\log\left(\int e^{\lambda\alpha(r-\int rdQ)}dQ\right)$$

we have by sub-Gaussianity:

$$\begin{aligned}\frac{1}{1-\alpha}\log\left(\int e^{\lambda(\alpha-1)(r-\int rdP)}dP\right) &\leq \frac{1}{1-\alpha}\frac{\lambda^2(1-\alpha)^2\sigma_P^2}{2} = \frac{\lambda^2(1-\alpha)\sigma_P^2}{2}\\ \frac{1}{\alpha}\log\left(\int e^{\lambda\alpha(r-\int rdQ)}dQ\right) &\leq \frac{1}{\alpha}\frac{\lambda^2\alpha^2\sigma_Q^2}{2} = \frac{\lambda^2\alpha\sigma_Q^2}{2}\end{aligned}$$

It follows that for all $\lambda\in(0,M)$

$$\begin{aligned}\lambda\left(\int rd\pi|x - \int rd\pi_{\text{ref}}|x\right) &\leq \frac{1}{\alpha}D_\alpha(\pi|x||\pi_{\text{ref}}|x) + \frac{\lambda^2(1-\alpha)\sigma_P^2}{2} + \frac{\lambda^2\alpha\sigma_Q^2}{2}\\ &= \frac{1}{\alpha}D_\alpha(\pi|x||\pi_{\text{ref}}|x) + \frac{\lambda^2((1-\alpha)\sigma_P^2+\alpha\sigma_Q^2)}{2}\end{aligned}$$

Integrating over x we obtain:

$$\lambda\left(\int rd\pi - \int rd\pi_{\text{ref}}\right) \leq \frac{1}{\alpha}D_{\alpha}(\pi||\pi_{\text{ref}}) + \frac{\lambda^2((1-\alpha)\sigma_P^2 + \alpha\sigma_Q^2)}{2}$$

Finally we have:

$$\int rd\pi - \int rd\pi_{\text{ref}} \leq \frac{1}{\lambda\alpha}D_{\alpha}(\pi||\pi_{\text{ref}}) + \frac{\lambda((1-\alpha)\sigma_P^2 + \alpha\sigma_Q^2)}{2}$$

minimizing over $\lambda \in (0, M)$: we obtain $\lambda^* = \sqrt{\frac{2D_{\alpha}(\pi||\pi_{\text{ref}})}{((1-\alpha)\sigma_P^2+\alpha\sigma_Q^2)\alpha}}$, M is free of choice, choosing $M = 2\lambda^*$, gives that λ^* is the minimizer and hence we have for all $\alpha \in (0,1)$:

$$\int rd\pi - \int rd\pi_{\text{ref}} \leq \sqrt{\frac{2((1-\alpha)\sigma_P^2 + \alpha\sigma_Q^2)D_{\alpha}(\pi||\pi_{\text{ref}})}{\alpha}}.$$

F Goodhart Laws

Proof 15 (Proof of Proposition 4) *We have by duality:*

$$\frac{1}{\beta}\log\left(\int e^{\beta r^*}d\pi_{\text{ref}}\right) = \sup_{\nu}\int r^*d\nu - \frac{1}{\beta}\mathsf{KL}(\nu||\pi_{\text{ref}})$$

hence for $\nu = \pi_{\beta,r}$ we have:

$$\frac{1}{\beta}\log\left(\int e^{\beta r^*}d\pi_{\text{ref}}\right) \geq \int r^*d\pi_{\beta,r} - \frac{1}{\beta}\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})$$

Hence:

$$\int r^*d\pi_{\beta,r} \leq \frac{1}{\beta}\log\left(\int e^{\beta r^*}d\pi_{\text{ref}}\right) + \frac{1}{\beta}\mathsf{KL}(\pi_{\beta,r}||\pi_{\text{ref}})$$

On the other hand by optimality of $\pi_{\beta,r}$ we have:

$$\mathsf{KL}\left(\pi_{\beta,r}||\pi_{\text{ref}}\right) = \beta\int rd\pi_{\beta,r} - \log\left(\int e^{\beta r}d\pi_{\text{ref}}\right)$$

Hence we have:

$$\int r^*d\pi_{\beta,r} \leq \frac{1}{\beta}\log\left(\int e^{\beta r^*}d\pi_{\text{ref}}\right) + \int rd\pi_{\beta,r} - \frac{1}{\beta}\log\left(\int e^{\beta r}d\pi_{\text{ref}}\right) \leq \int rd\pi_{\beta,r} + \frac{1}{\beta}\log\left(\frac{\int e^{\beta r^*}d\pi_{\text{ref}}}{\int e^{\beta r}d\pi_{\text{ref}}}\right)$$

It follows that:

$$\int r^*d\pi_{\beta,r} - \int r^*d\pi_{\text{ref}} \leq \int rd\pi_{\beta,r} - \int rd\pi_{\text{ref}} + \frac{1}{\beta}\log\left(\frac{\int e^{\beta(r^* - \int r^*d\pi_{\text{ref}})}d\pi_{\text{ref}}}{\int e^{\beta(r - \int rd\pi_{\text{ref}})}d\pi_{\text{ref}}}\right)$$

$$\begin{aligned}
\frac{\int e^{\beta(r^* - \int r^*d\pi_{\text{ref}})}d\pi_{\text{ref}}}{\int e^{\beta(r - \int rd\pi_{\text{ref}})}d\pi_{\text{ref}}} &= \int e^{\beta(r^* - r - (\int r^*d\pi_{\text{ref}} - \int rd\pi_{\text{ref}})}\frac{e^{\beta r}d\pi_{\text{ref}}}{\int e^{\beta r}d\pi_{\text{ref}}} \\
&= \int e^{\beta(r^* - r - (\int r^*d\pi_{\text{ref}} - \int rd\pi_{\text{ref}})}d\pi_{\beta,r}
\end{aligned}$$

Hence we have finally:

$$\int r^* d\pi_{\beta,r} - \int r^* d\pi_{\text{ref}} \leq \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} + \frac{1}{\beta} \log \left(\int e^{\beta(r^* - r - \left(\int r^* d\pi_{\text{ref}} - \int r d\pi_{\text{ref}}\right)} d\pi_{\beta,r} \right)$$

$$\int r^* d\pi_{\beta,r} - \int r^* d\pi_{\text{ref}} \leq \int r d\pi_{\beta,r} - \int r d\pi_{\text{ref}} - \frac{1}{\beta} \log \left(\int e^{\beta(r - r^* - \left(\int r d\pi_{\text{ref}} - \int r^* d\pi_{\text{ref}}\right)} d\pi_{\beta,r^*} \right)$$

The proof follows from using the subgaussianity of $r_\sharp \pi_{\text{ref}}$ *and the assumption on the soft max.*

Proof 16 (Proof of Proposition 5)

$$\mathbb{E}_{\pi}(r^* - r) - \mathbb{E}_{\pi_{\text{ref}}}(r^* - r) \leq 2||r - r^*||_{\infty} \mathsf{TV}(\pi, \pi_{\text{ref}})$$

For $\pi^{(n)}_{r,\text{ref}}$*, we have:*

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \leq \mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r) - \mathbb{E}_{\pi_{\text{ref}}}(r) + 2||r - r^*||_{\infty} \mathsf{TV}(\pi^{(n)}_{r,\text{ref}}, \pi_{\text{ref}})$$

and

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \geq \mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r) - \mathbb{E}_{\pi_{\text{ref}}}(r) - 2||r - r^*||_{\infty} \mathsf{TV}(\pi^{(n)}_{r,\text{ref}}, \pi_{\text{ref}})$$

By the data processing inequality we have: $\mathsf{TV}(\pi^{(n)}_{r,\text{ref}}, \pi_{\text{ref}}) \leq \mathsf{TV}(R^{(n)}_{r,\text{ref}}, R) = (\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}}$ *If* r *has subguassian tails under* π_{ref} *than we have:*

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \leq \sqrt{2\sigma^2 \left(\log(n) - \frac{n-1}{n} \right)} + 2||r - r^*||_{\infty} \left((\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}} \right)$$

$$\mathbb{E}_{\pi^{(n)}_{r,\text{ref}}}(r^*) - \mathbb{E}_{\pi_{\text{ref}}}(r^*) \leq \sqrt{2\sigma^2 \left(\log(n) - \frac{n-1}{n} \right)} + 2 \inf_{r \in \mathcal{H}} ||r - r^*||_{\infty} \left((\frac{1}{n})^{\frac{1}{n-1}} - (\frac{1}{n})^{\frac{n}{n-1}} \right).$$

G Supplementary Figures and Experiments

G.1 Tails of Reward model FsfairX-LLaMA3-RM-v0.1 evaluated on Popular LLMs

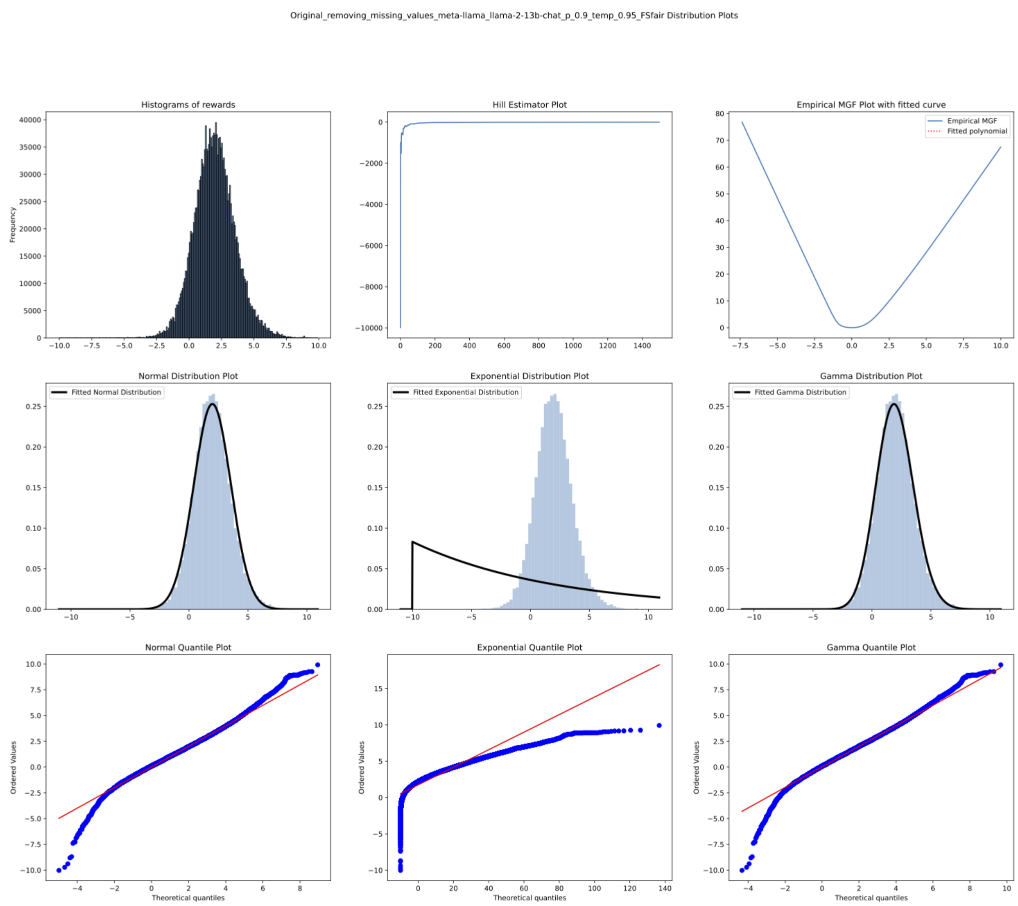

Figure 4: Reward evaluated on LLama2-7B. We see that the reward follows more a Gaussian or a gamma random variable and it is not heavy tailed. The Momemt generating function (MGF) follows a quadratic The Hill index is not meaningful in this case.

G.2 Reward versus KL on other LLMs with best of n policies

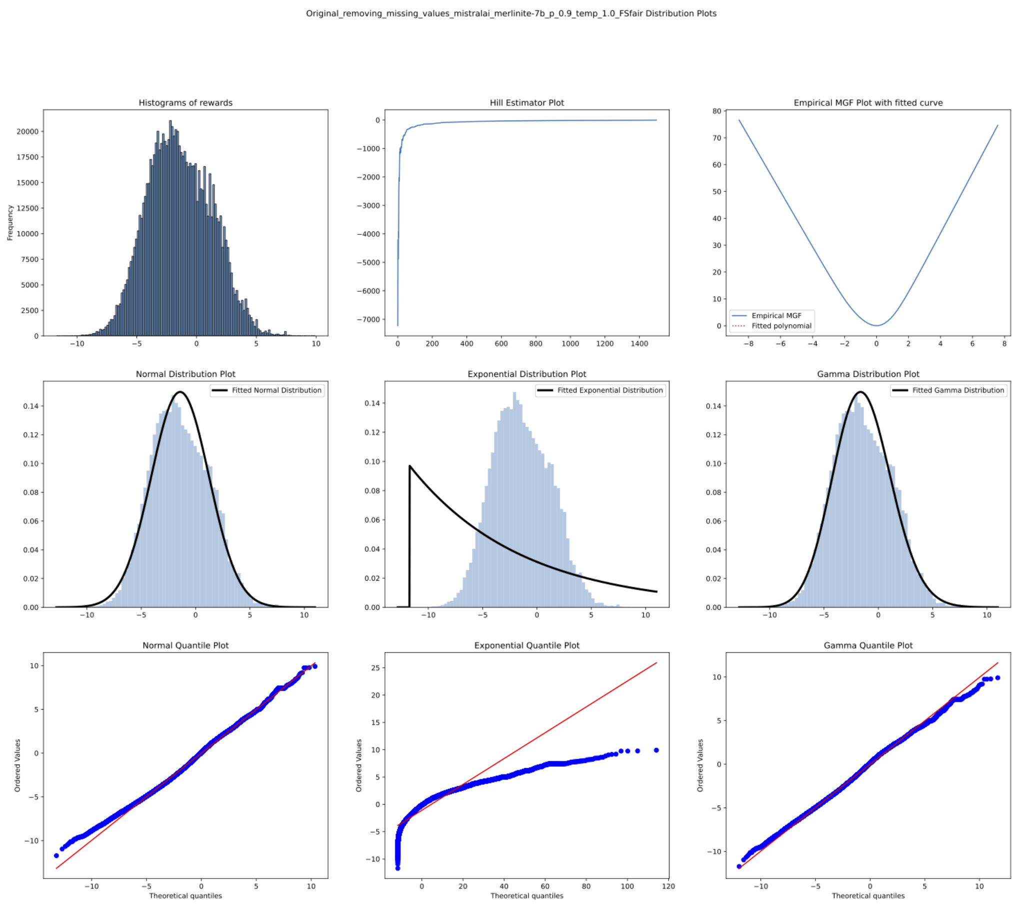

Figure 5: Reward evaluated on Merlinite 7B. The reward follows a gamma distribution as it is almostly perfectly matching it in the q-q plots (quantile plots).

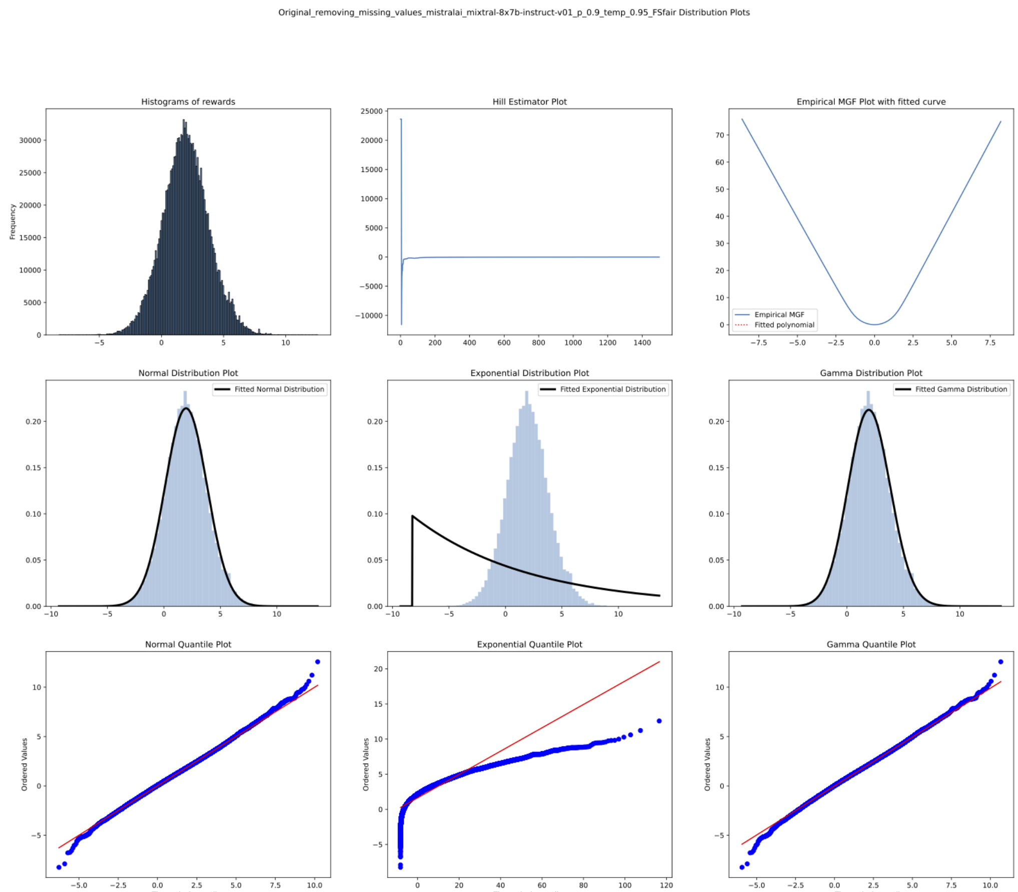

Figure 6: Reward evaluated on Mixtral8x7b. The reward follows a gaussian or a gamma distribution and is not heavy tailed.

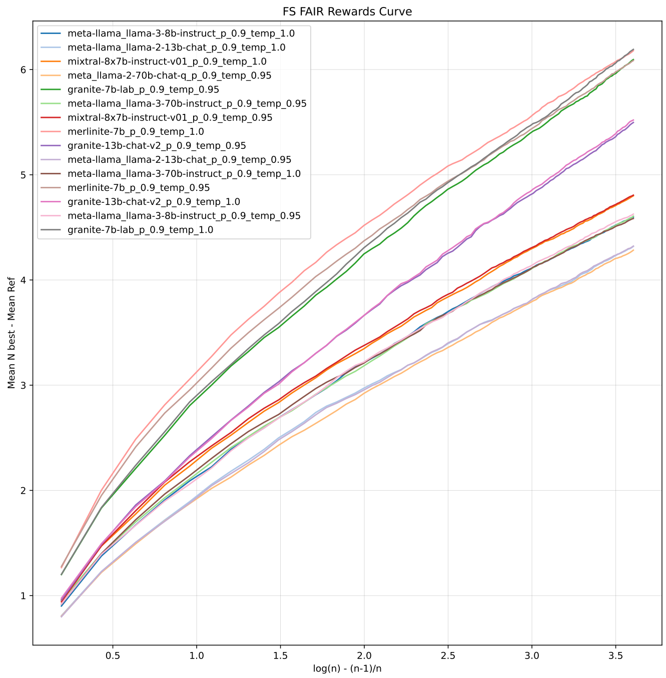

Figure 7: Centered Reward (FsfairX-LLaMA3-RM-v0.1) versus KL of best of n policy for various LLM

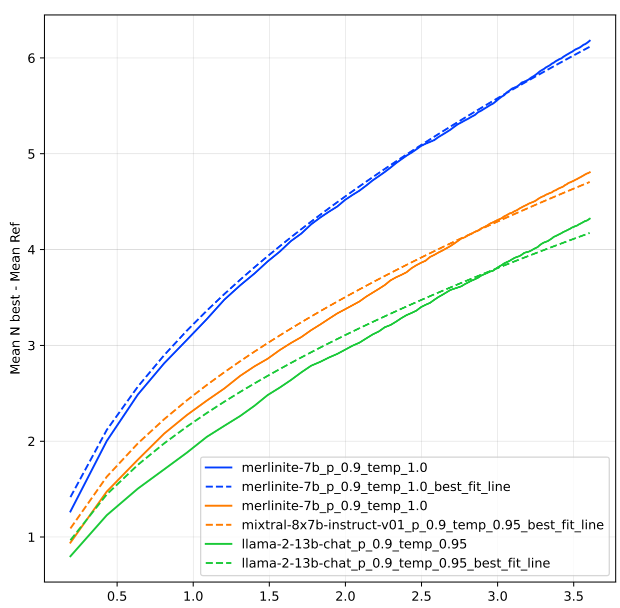

Figure 8: Best fit for $y = a * \sqrt{x}$ for centered best of n reward versus KL. We see that this fit is not as good as $y = a * \sqrt{x} + b * x$, hitting to a subgamma tail rather than subgaussian.

H Experiments with OAS Reward Model OpenAssistant/reward-model-deberta-v3-large-v2

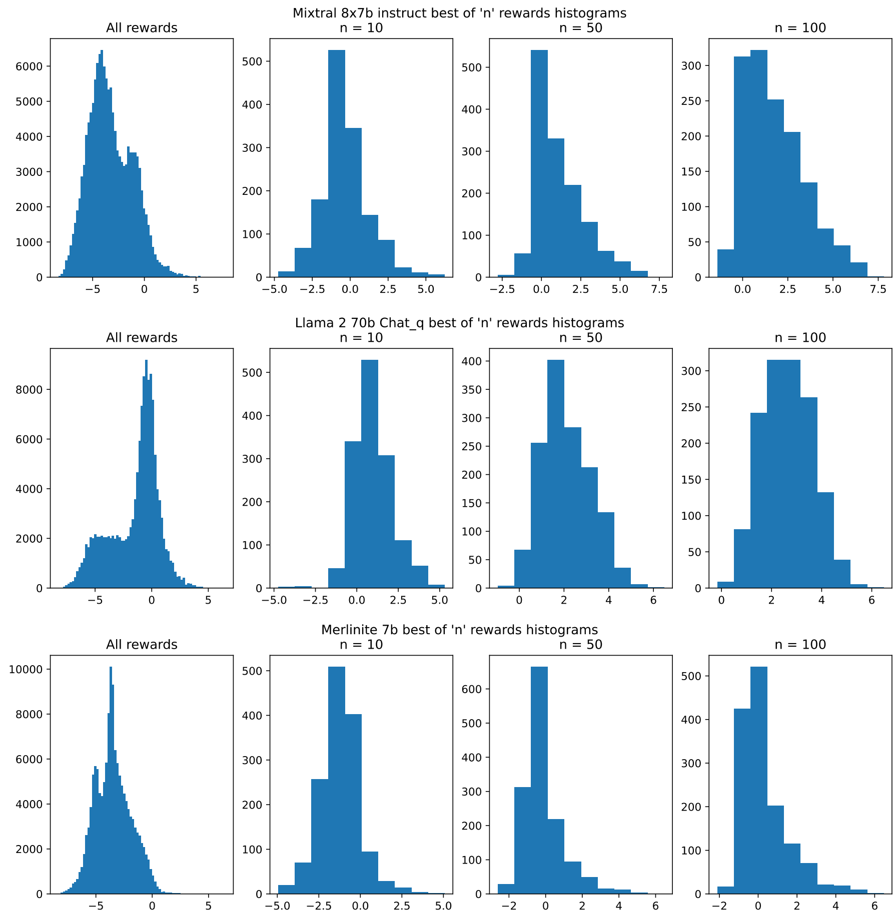

Figure 9: Histograms of OAS Reward for reference and best of n policies.

Model and n	Mean	Std	Max
Mixtral all	-3.25	2.11	7.8
Mixtral n =10	-0.34	1.5	6.23
Mixtral n =50	1.1	1.57	7.8
Mixtral n =100	1.76	1.64	7.8
Llama2_70b_chat all	-1.47	2.17	6.48
Llama2_70b n =10	0.95	1.15	5.29
Llama2_70b n =50	2.12	1.09	6.48
Llama2_70b n =100	2.63	1.01	6.48
Merlinite all	-3.47	1.6	6.46
Merlinite n =10	-1.23	1.19	5.07
Merlinite n =50	-0.14	1.08	6.46
Merlinite n =100	0.31	1.2	6.46

Table 4: Statistics of OpenAssistant/reward-model-deberta-v3-large-v2 reward evaluated for reference and best of n policies for different n values.

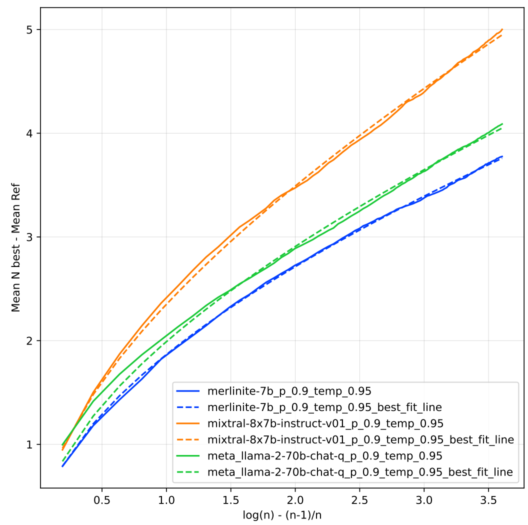

Figure 10: Centered Reward (OAS) versus KL best of n policies, with best fit $y = a\sqrt{x} + bx$. The fit hints to subgamma tails.

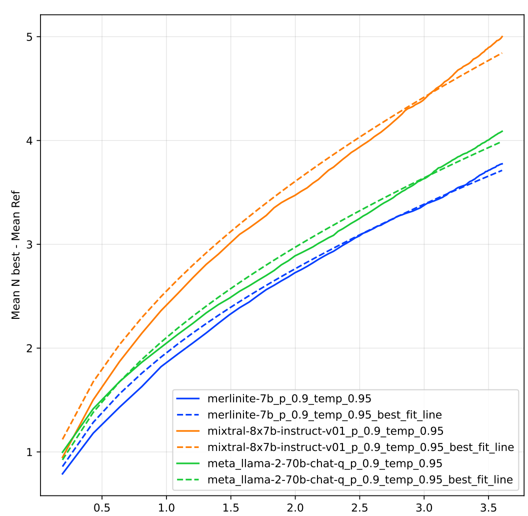

Figure 11: Centered Reward (OAS) versus KL best of n policies, with best fit $y = a\sqrt{x}$. This subgaussian fit provides an upper bound that is not as tight as above.

Model	a	b
Merlinite-7B	1.74119974	0.12555422
Mixtral-8x7B	2.06674977	0.28401807
Llama2-70B chat	1.83327634	0.15764552

Table 5: OAS Reward model: $y = a\sqrt{x} + bx$ best fitted coefficients for centered reward versus KL. a is slightly larger than std.

Model	a
Merlinite-7B	1.9550695
Mixtral-8x7B	2.55054775
Llama-70B-Chat	2.10181062

Table 6: OAS Reward model: $y = a\sqrt{x}$ best fitted coefficients for centered reward versus KL.

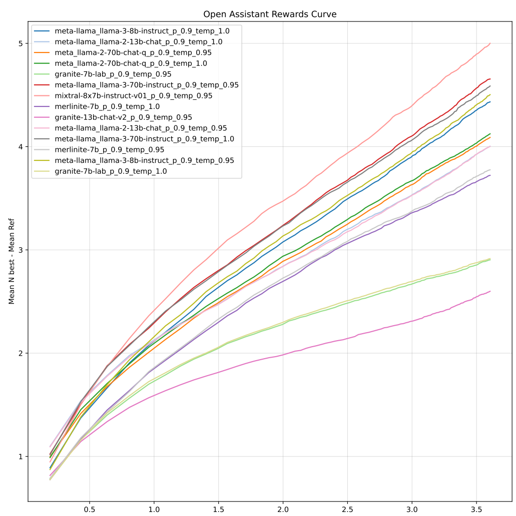

Figure 12: OAS Reward: Centered Rewards versus KL best of n policies for various models.

I Reward Versus Rényi In Best of N

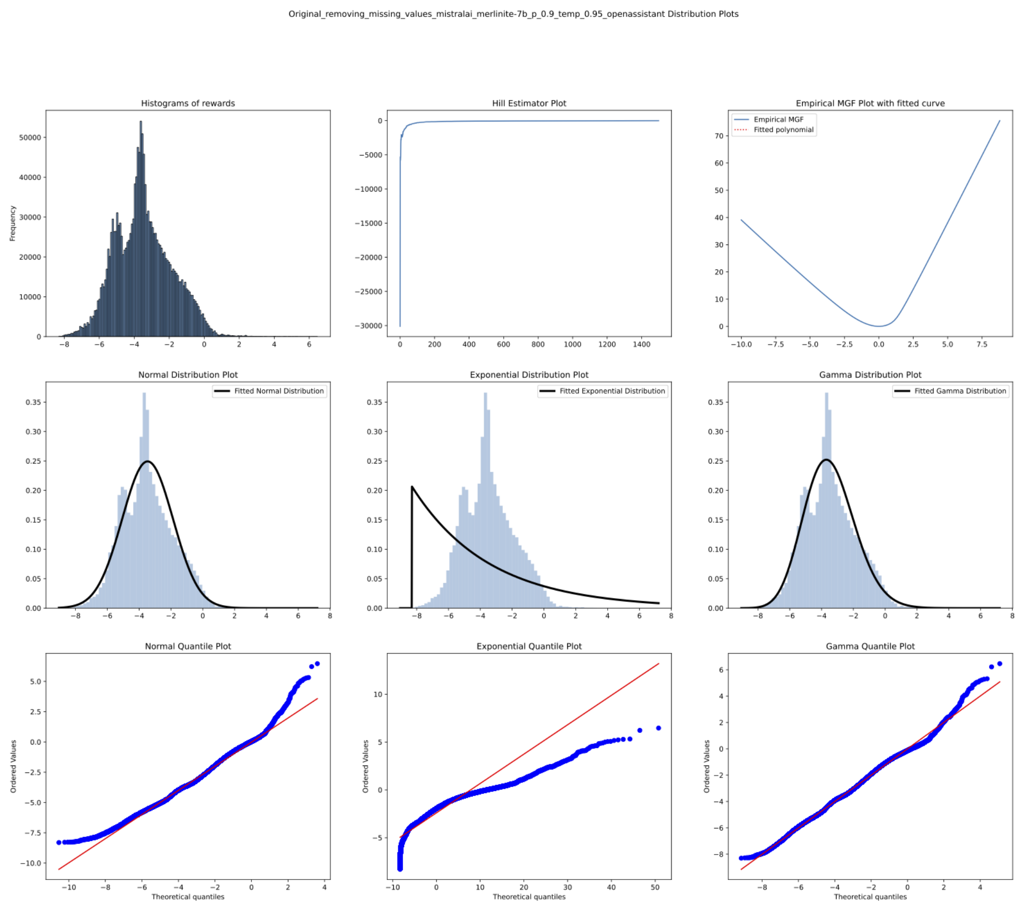

Figure 13: subgamma tails of the OAS reward for Merlinite 7B as seen in the q-q plots.

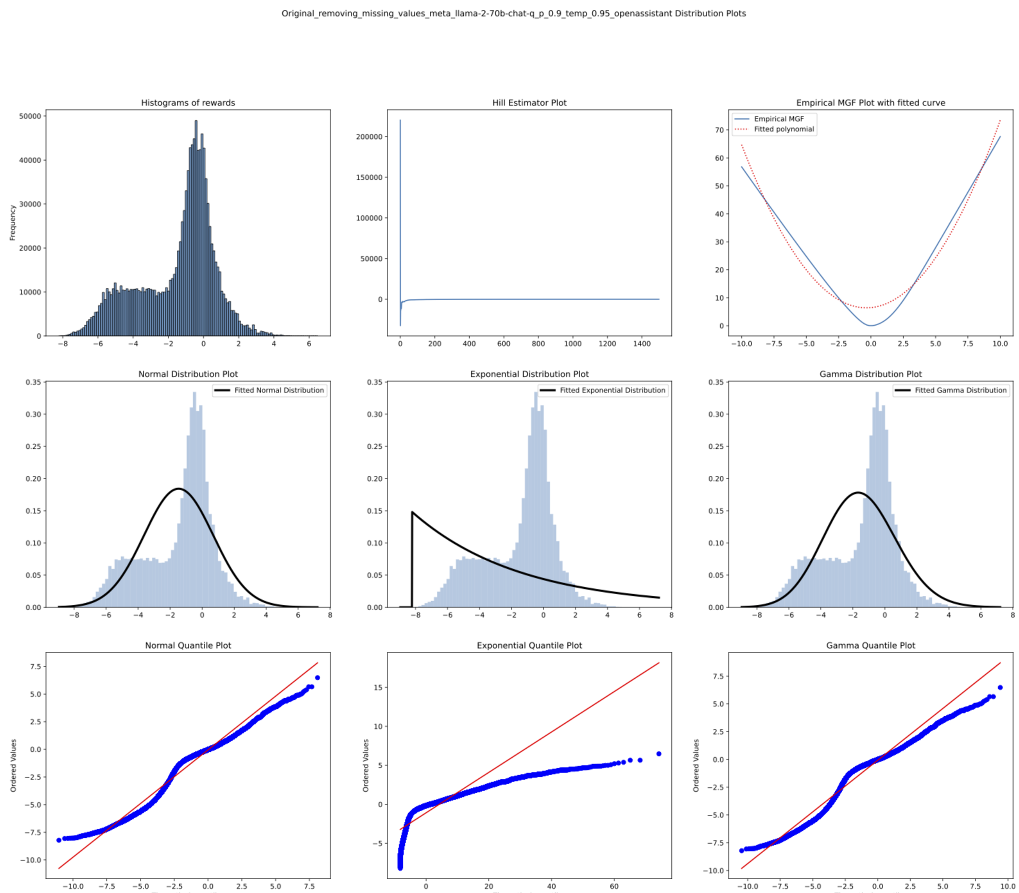

Figure 14: subgaussian/subgamma tails of the OAS reward for LLama2-70B-chat.

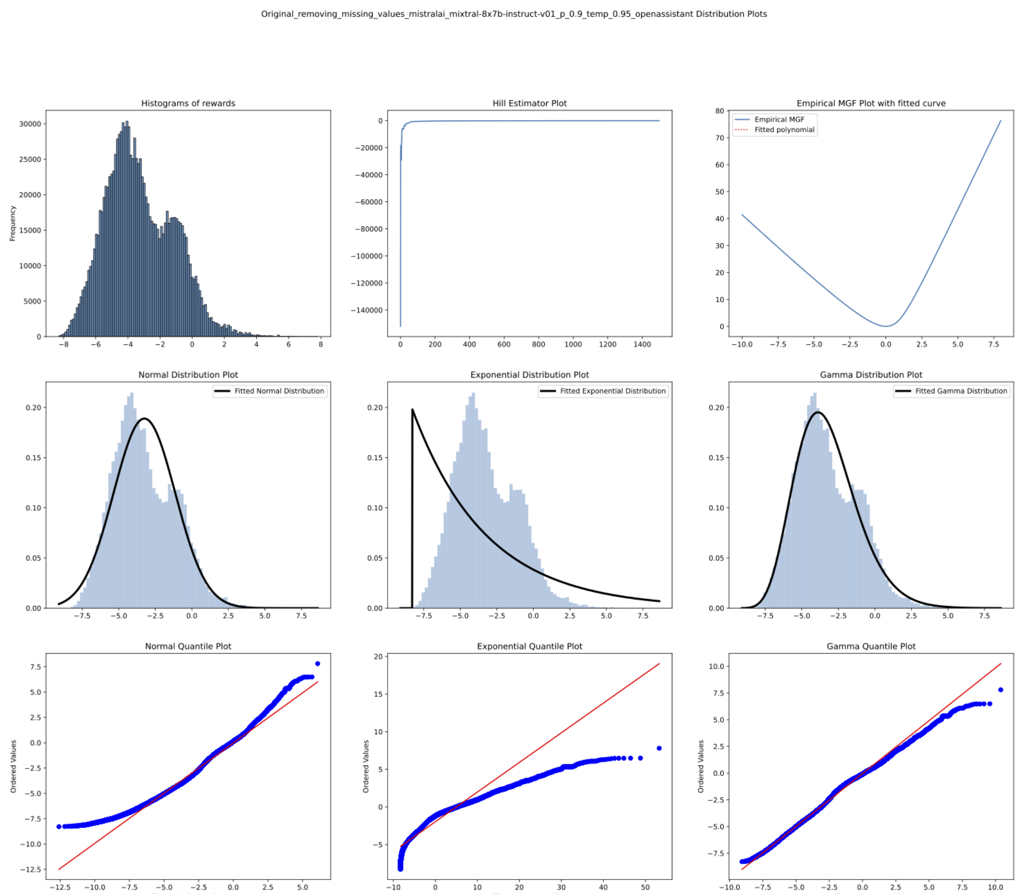

Figure 15: Subgamma tails of the OAS reward of Mixtral-8x7b-instruct.

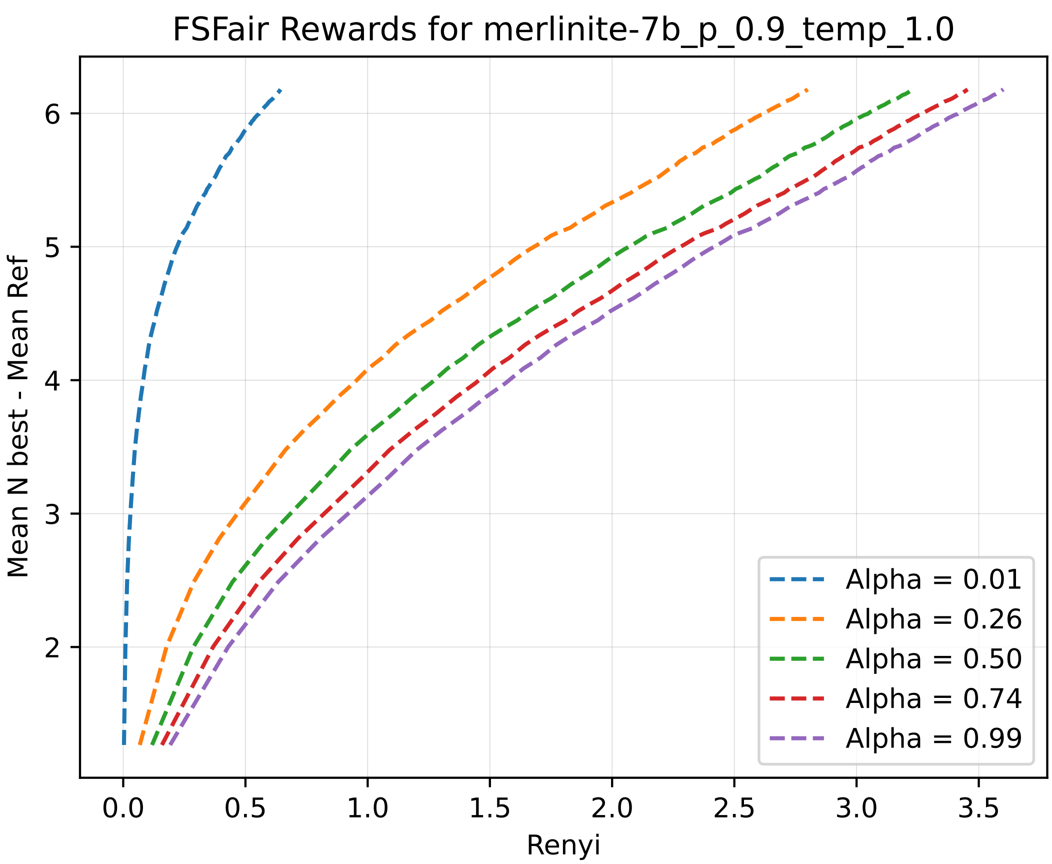

Figure 16: Centered Reward for Best of N versus Renyi divergence for $\alpha \in (0, 1)$ - Merlinite

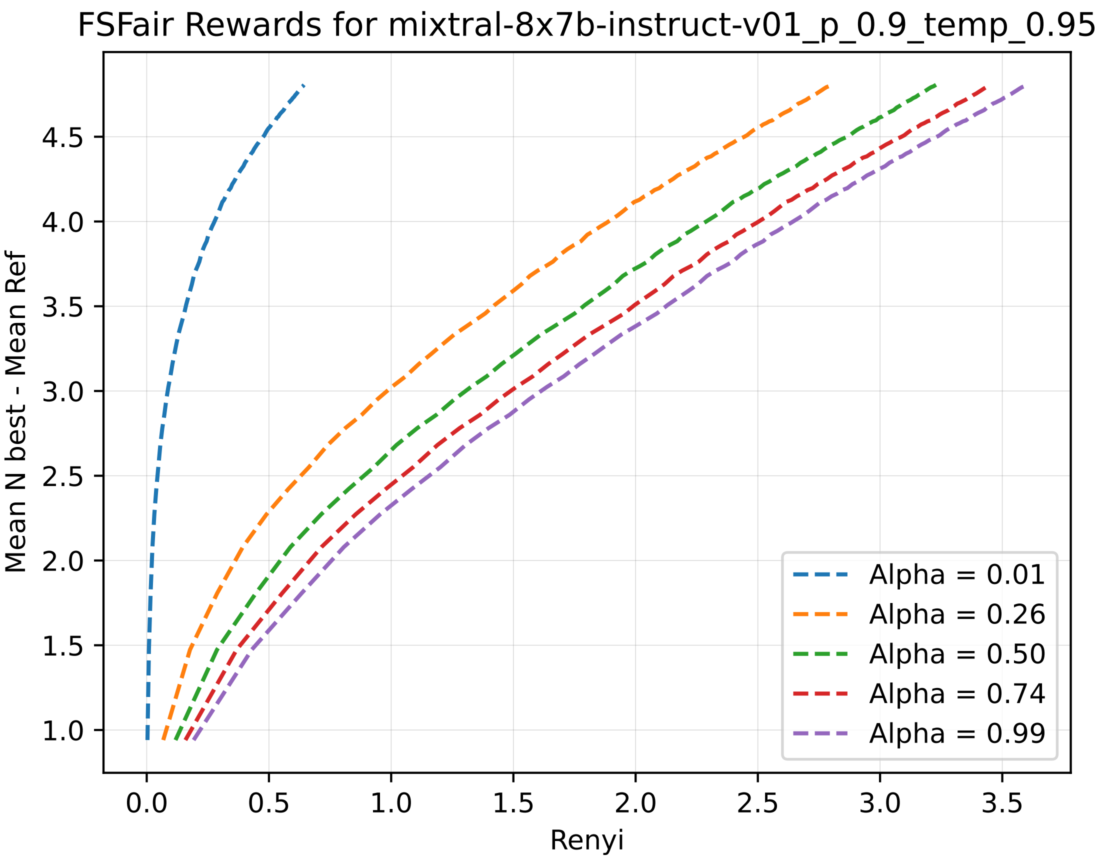

Figure 17: Centered Reward for Best of N versus Renyi divergence for $\alpha \in (0, 1)$ -Mixtral

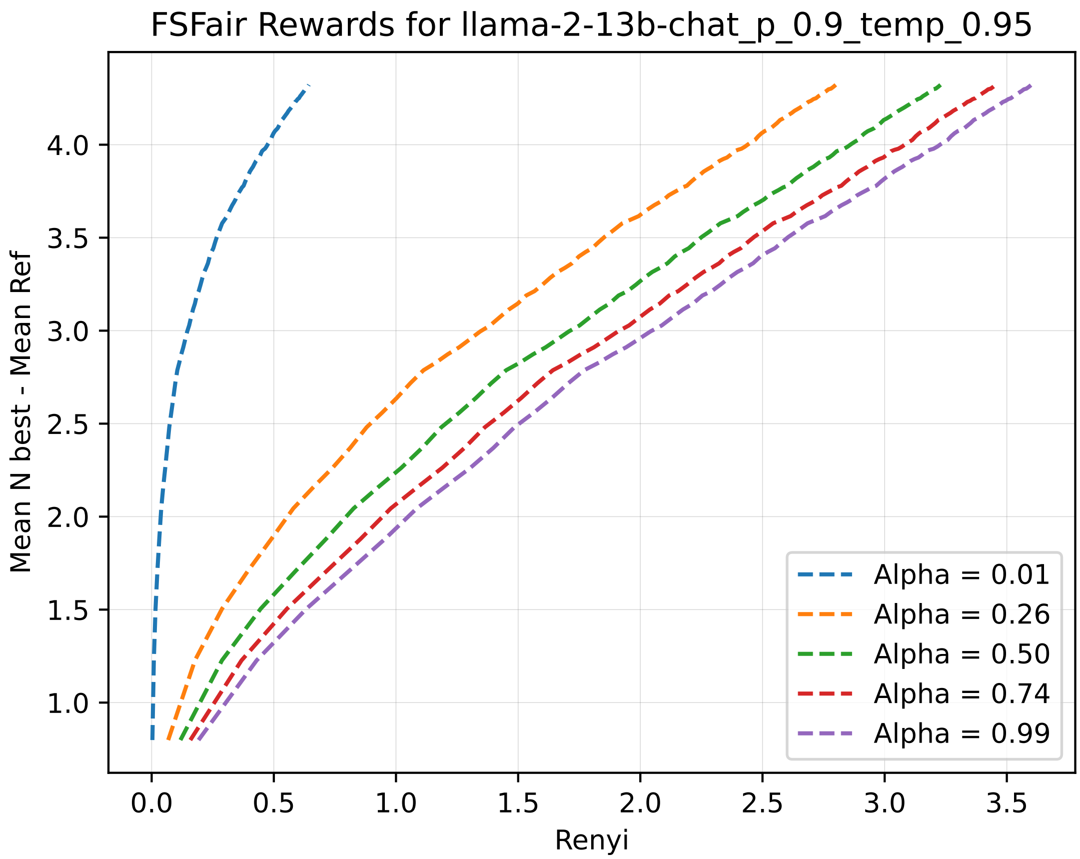

Figure 18: Centered Reward for Best of N versus Renyi divergence for $\alpha \in (0, 1)$ -LLama

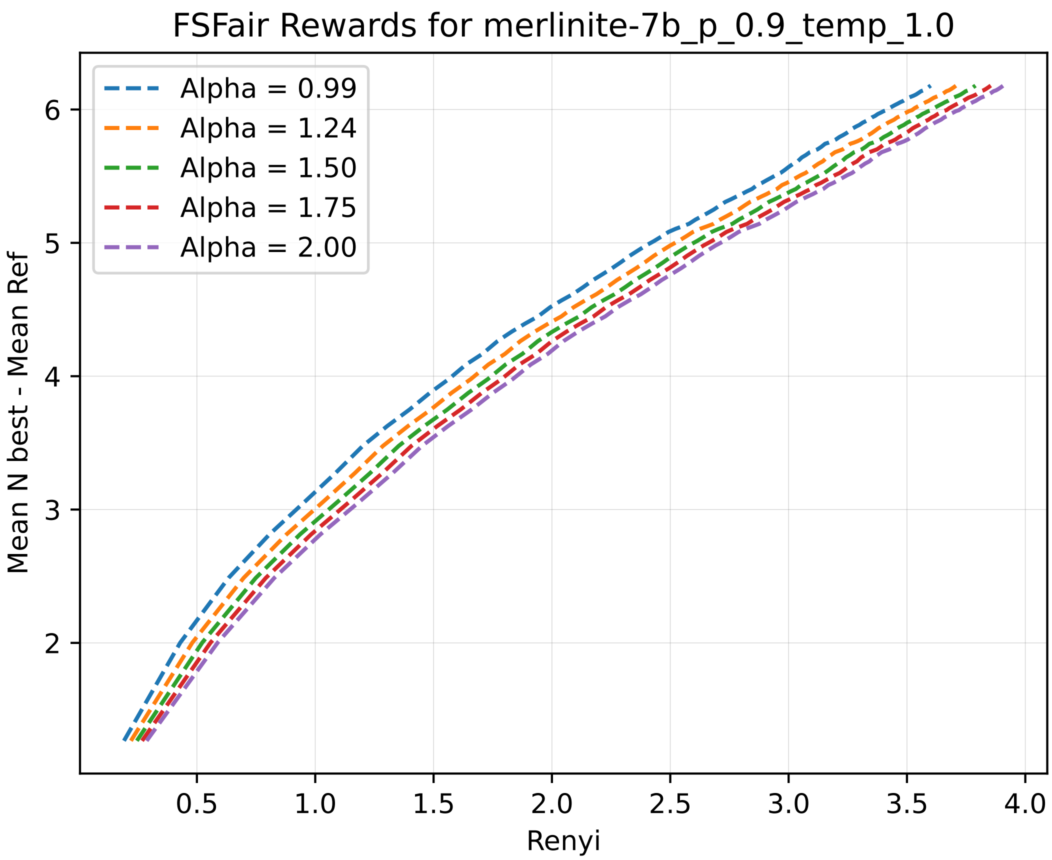

Figure 19: Centered Reward for Best of N versus Renyi divergence for $\alpha > 1$ - Merlinite

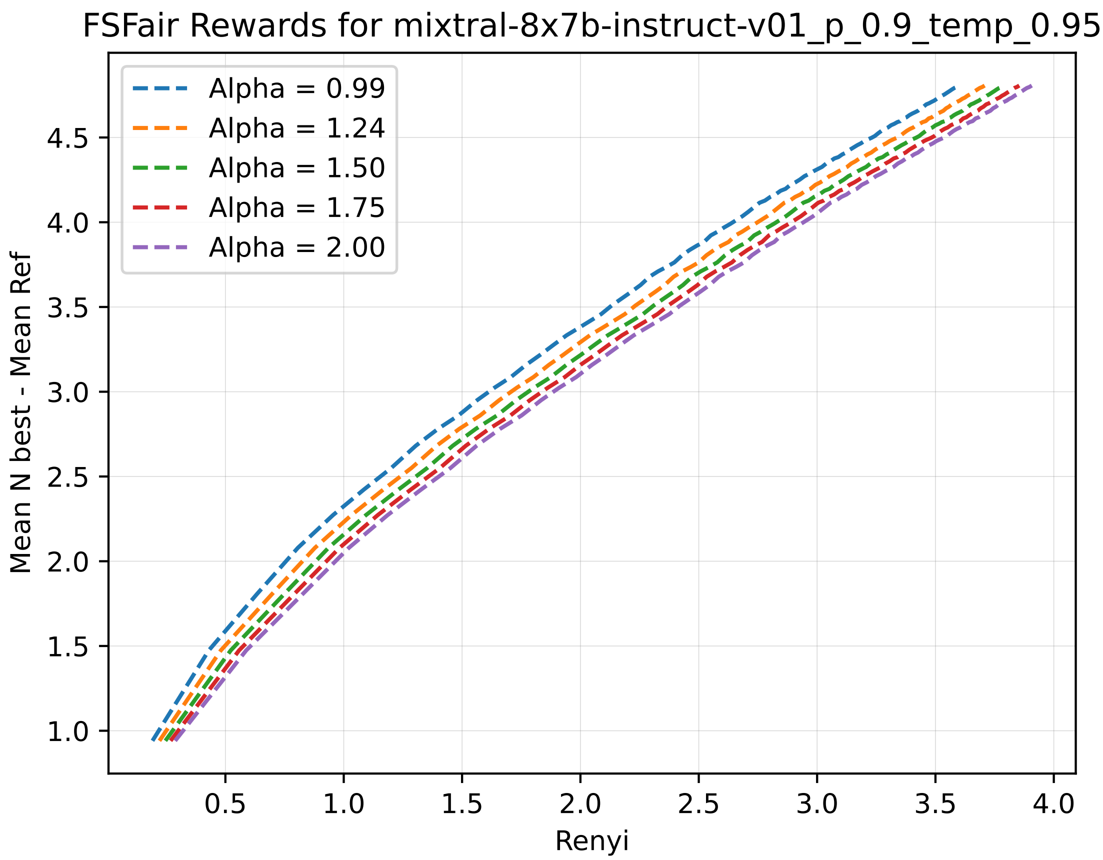

Figure 20: Centered Reward for Best of N versus Renyi divergence for $\alpha > 1$ -Mixtral

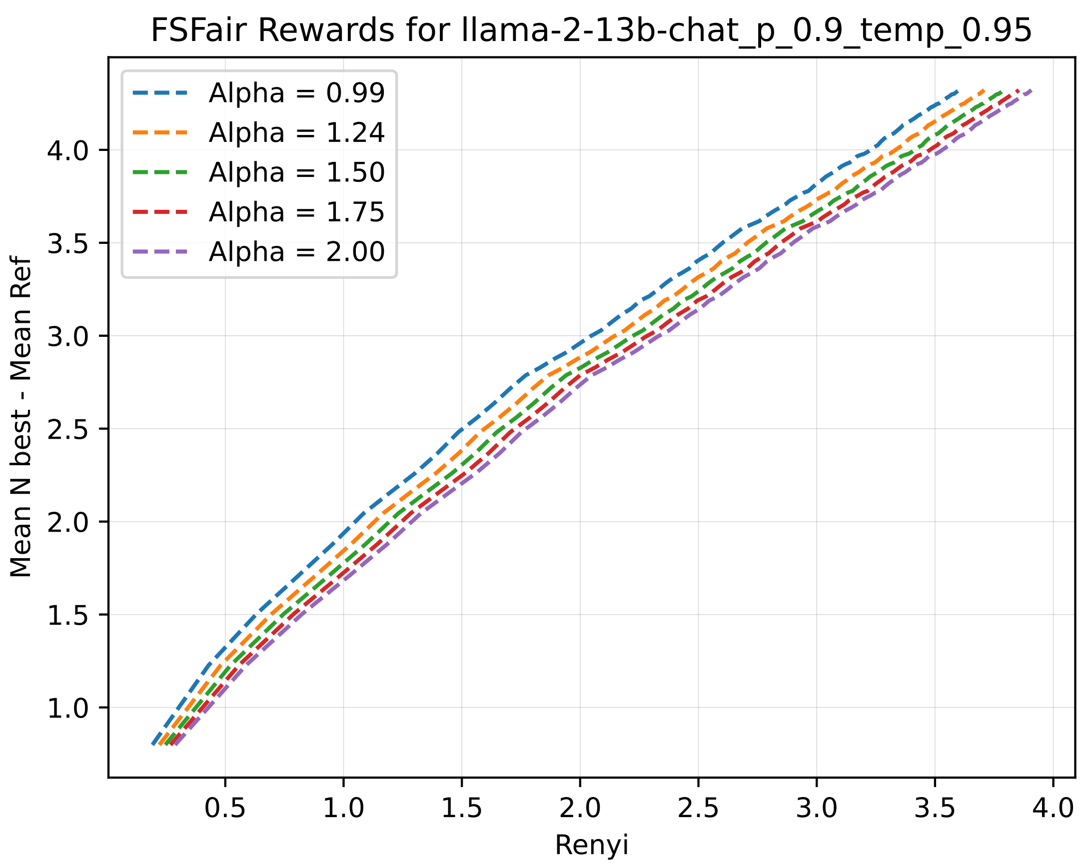

Figure 21: Centered Reward for Best of N versus Renyi divergence for $\alpha > 1$ -LLama