BEST-OF-N THROUGH THE SMOOTHING LENS: KL DIVERGENCE AND REGRET ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

A simple yet effective method for inference-time alignment of generative models is Best-of-N (BoN), where N outcomes are sampled from a reference policy, evaluated using a calibrated proxy-reward model, and the highest-scoring one is selected. While prior work argues that BoN is almost optimal in reward vs KL tradeoffs, the effectiveness of BoN depends critically on the quality of the (calibrated) proxy-reward model used for selection. For this purpose, we study BoN through a smooth version known as Soft Best-of-N (SBoN) and develop a theoretical framework to address this gap. We analyze the scaling behaviour of BoN by providing bounds on the KL divergence between the SBoN policy and the reference policy, offering insights into how performance varies with the number of samples. We also study the regret gap, i.e., the gap between the expected calibrated true-reward under the optimal policy and the SBoN policy. Our theoretical and empirical findings show that smoothing helps SBoN mitigate reward overoptimization, especially when the quality of the calibrated proxy-reward is low.

1 INTRODUCTION

Large language models (LLMs) have transformed machine learning, achieving state-of-the-art results on a variety of tasks. Despite all advancements, LLMs can still generate undesirable outputs, such as toxic or factually incorrect responses. This has made alignment a central goal in modern LLM development (Achiam et al., 2023; Team et al., 2023).

Several post-hoc alignment methods have been proposed to address this challenge, including Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022), SLiC (Zhao et al., 2022), Direct Preference Optimization (Rafailov et al., 2023), controlled decoding (Mudgal et al., 2024) and Best-of-N (BoN) sampling (Beirami et al., 2024). While these methods differ in their implementationranging from training-time optimization to test-time selectionthey can be viewed, in principle, as approximating the solution to a KL-regularized reward maximization problem. The optimal solution to this problem is a tilted distribution over responses, which balances reward and proximity to the reference model (Yang et al., 2024).

In BoN as a test-time sampling strategy, given a prompt, N responses are sampled from the reference policy, and the one with the highest proxy-reward sample is selected. Empirically, BoN has been shown to achieve competitive or superior performance in the reward-versus-KL divergence trade-off when compared to RLHF and other alignment methods (Gao et al., 2023a; Mudgal et al., 2024) under *true-reward model*. Furthermore, under certain conditions, it asymptotically approximates the solution to the KL-regularized reward maximization objective (Yang et al., 2024). However, in practice, BoN relies on a learned proxy-reward modelan approximation of the true-reward functionto guide this selection. As such, their effectiveness critically depends on both the *proxy-reward model* (estimation error of true-reward) and the *quality of the reference policy*.

Understanding how these two components, the quality of the calibrated proxy-reward model and the choice of reference policy, affect the alignment quality of test-time sampling algorithms is essential. There are different measures of alignment quality, including KL divergence [1] between aligned policy and reference policy and the *regret* defined as the gap between the expected calibrated true-reward

[1] Unless stated otherwise, all KL divergences are understood to be measured between the aligned policy and the reference policy.

under the optimal (tilted) policy and the alignment policy. Note that minimizing the regret gap is critical to ensuring high-quality outputs and close performance to the optimal policy. Recent work by Gao et al. (2023a) and Hilton et al. (2022) has investigated the scaling laws governing reward model optimization in both reinforcement learning (RL) and BoN settings as a function of KL divergence between aligned policy and reference policy. They empirically demonstrate that, under proxy-reward models, the improvement in expected calibrated true-reward, relative to a reference policy, scales proportionally for both RL and BoN policies.

While recent work analyzes BoN under the idealized settings where there is no discrepancy between the calibrated proxy-reward and the true-reward (Yang et al., 2024; Beirami et al., 2024; Mroueh, 2024; Huang et al., 2025), our work relaxes this assumption to study the interplay between the reward discrepancy measured through regret and the KL-divergence. We present a theoretical study of **Soft Best-of-N (SBoN)**, a smoothed variant of BoN recently introduced by Mayrink Verdun et al. (2025); Jinnai et al. (2024). Unlike BoN, SBoN draws the final response *probabilistically* from the N candidates, yielding a policy that is tunable with a temperature parameter. Our analysis centres on two metrics:

(a) the KullbackLeibler divergence between SBoN policy (under the *true* reward or proxy-reward model) and the reference policy, and
(b) the *regret*, i.e. the expected calibrated true-reward gap between optimal policy and SBoN policy.

We show how these results specialize to the BoN (as a limit of SBoN for the temperature goes to infinity) and quantify the estimation error incurred by using a *proxy* reward model instead of the true-reward. Finally, we characterize regimes in which SBoN attains *lower* regret bound than BoN when we use the calibrated proxy-reward model. Our main contributions are:

- We derive finite-sample bounds for KL divergence between the SBoN policy and reference policy, and for the regret gap of the SBoN policy, and we extend these bounds to BoN. These bounds reveal how the number of responses N, (calibrated) proxy-reward model quality and reference policy model affect performance.
- We quantify cases where SBoN performs better than BoN under overoptimization scenario where the calibrated proxy-reward model is used instead of the calibrated true-reward model.
- We provide experimental validation using various proxy-reward models to demonstrate SBoN's advantages in the overoptimization scenario.

2 Related Works

In this section, we discuss related works on BoN, the theoretical foundation of (Soft) BoN and overoptimization. More related works for the theoretical foundation of RLHF and smoothing of maximum are provided in the Appendix (App) A.

Best-of-N: Despite many recent advancements in alignment, a simple, popular, and well-performing method continues to be the BoN policy (Nakano et al., 2021; Stiennon et al., 2020; Beirami et al., 2024). In fact, Gao et al. (2023b); Mudgal et al. (2024); Eisenstein et al. (2023) show that BoN consistently achieves compelling win rateKL tradeoff curves, often outperforming KL-regularized reinforcement learning and other more complex alignment strategies. LLaMA 2 (Touvron et al., 2023) leverages BoN outputs as teacher signals to further finetune the base model. Mudgal et al. (2024) extend BoN through Q-learning to block-wise BoN decoding. This empirical effectiveness has also inspired research into distilling BoN behaviour into standalone models (Amini et al., 2025; Sessa et al., 2024; Qiu et al., 2024). Hughes et al. (2024) utilize BoN as an effective method for jailbreaking, while BoN is also commonly used as a strong baseline for scaling inference-time compute (Brown et al., 2024; Snell et al., 2024). Given the broad success of BoN, we are motivated to theoretically investigate the BoN policies and the effect of the calibrated proxy-reward model (reward hacking) and the quality of the reference policy.

Theoretical Foundation of (Soft) BoN: KL divergence of BoN is studied in (Beirami et al., 2024; Mroueh, 2024) via information theoretical tools where the KL divergence of BoN sampling from the reference distribution is bounded by $\log(N) - (N-1)/N$. Scaling laws governing reward as a function of KL divergence is empirically studied by Gao et al. (2023b) and theoretically formalized by Mroueh (2024). Furthermore, the asymptotic case and the equivalence of BoN to the KL-constrained reinforcement learning solution are studied by Yang et al. (2024) under the assumption of access to optimal reward. Gui et al. (2024) further characterized the win rateKL gap in the asymptotic regime where a model assigns extremely low likelihoods to successful completions. Furthermore, Sun et al. (2024) accelerated BoN using speculative rejection sampling. The regret of BoN under some assumptions is studied in (Huang et al., 2025). The convergence rate of the SBoN policy to the optimal tilted policy has been analyzed by Mayrink Verdun et al. (2025). Additionally, Geuter et al. (2025) investigate a variant of SBoN that incorporates speculative samples from a small auxiliary model, providing both theoretical and empirical insights. However, the regret gap and KL divergence of SBoN under overoptimization scenario remain largely unexplored in the existing literature.

3 Problem Formulation

Notations: Upper-case letters denote random variables (e.g., Z), lower-case letters denote the realizations of random variables (e.g., z), and calligraphic letters denote sets (e.g., $\mathcal{Z}$). All logarithms are in the natural base. The set of probability distributions (measures) over a space $\mathcal{X}$ with finite variance is denoted by $\mathcal{P}(\mathcal{X})$. Δ_N is N-simplex distribution set. The KL divergence between two probability distributions on $\mathbb{R}^d$ with densities $p(x)$ and $q(x)$, such that $q(x) > 0$ when $p(x) > 0$, is $\mathrm{KL}(p\|q) := \int_{\mathbb{R}^d} p(x)\log(p(x)/q(x))\mathrm{d}x$ (with $0.\log 0 := 0$). The total-variation distance is defined as $\mathbb{TV}(p,q) = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)|$. Furthermore, we define chi-square divergence as $\chi^2(p(x)\|q(x)) = \int_{\mathcal{X}} \frac{p^2(x)}{q(x)} - 1$.

3.1 Preliminaries

Let the finite set of prompts be $\mathcal{X}$ and the discrete finite set of responses be $\mathcal{Y}$. Prompts are drawn from a distribution ρ over $\mathcal{X}$. A (stochastic) policy $\pi \in \Pi$ assigns, for every prompt $x \in \mathcal{X}$, a conditional distribution $\pi(\cdot \mid x)$ over $\mathcal{Y}$; drawing $y \sim \pi(\cdot \mid x)$ yields a response.

We treat the supervised-fine-tuned (SFT) model as a *reference policy*, denoted $\pi_{\mathrm{ref}}(\cdot \mid x)$. We consider a *calibrated* true-reward function, (Balashankar et al., 2025), $r^\star_{\mathrm{c}} : \mathcal{X} \times \mathcal{Y} \to [0,1]$. For a given temperature $\beta > 0$, we seek a policy that remains close to π_{ref} while maximizing expected calibrated true-reward, leading to the KL-regularized objective

$$\max_{\pi\in\Pi} \mathbb{E}_{Y\sim\pi(\cdot|x)}\big[r^\star_{\mathrm{c}}(x,Y)\big] \;-\; \frac{1}{\beta}\,\mathrm{KL}\big(\pi(\cdot\mid x)\,\|\,\pi_{\mathrm{ref}}(\cdot\mid x)\big). \tag{1}$$

The unique solution is the *tilted* optimal policy (Korbak et al., 2022b;a; Yang et al., 2024)

$$\pi_{\beta,r^\star_{\mathrm{c}}}(y|x) \;=\; \frac{\pi_{\mathrm{ref}}(y\mid x)\,\exp\big(\beta\, r^\star_{\mathrm{c}}(x,y)\big)}{Z_{r^\star_{\mathrm{c}},Y}(x,\beta)}, \tag{2}$$

where $Z_{r^\star,Y}(x,\beta) = \sum_{y\in\mathcal{Y}} \pi_{\mathrm{ref}}(y\mid x)\,\exp\big(\beta\, r^\star_{\mathrm{c}}(x,y)\big)$, is the normalizing (*partition*) function.

Note that, in practice, we do not have access to the closed form of reference policy $\pi_{\mathrm{ref}}(y|x)$ and $r^\star_{\mathrm{c}}(y,x)$. We can only first estimate the true-reward function via *calibrated* proxy-reward function $\hat{r}_{\mathrm{c}}(y,x)$ using some datasets and then sample from $\pi_{\mathrm{ref}}(y|x)$ and compute $\hat{r}_{\mathrm{c}}(y,x)$ for each individual sample. Then, we can apply inference time algorithms, e.g., BoN or SBoN (Mayrink Verdun et al., 2025), where N samples are generated from $\pi_{\mathrm{ref}}(y|x)$ and we choose the sample with the highest calibrated proxy-reward function (BoN) or sampled from a distribution (SBoN) using the calibrated proxy-reward function. When only a calibrated proxy-reward function $\hat{r}_{\mathrm{c}}(y,x)$ is available, we obtain the analogous partition function $Z_{\hat{r}_{\mathrm{c}},Y}(x,\beta)$ and policy $\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)$.

3.2 Calibrated Reward Function

In this work, we consider calibrated reward functions instead of raw (uncalibrated) reward functions. As shown by Balashankar et al. (2025), calibrated reward function satisfies:

- Boundedness: for all x, y, we have $\hat{r}_{\mathrm{c}}(y,x),\ r^\star_{\mathrm{c}}(y,x) \in [0,1]$.

- Uniformity under the reference model: for each prompt $x \in \mathcal{X}$, if $y \sim \pi_{\text{ref}}(\cdot \mid x)$ then $r_{\text{c}}(y, x) \sim \text{Unif}(0, 1)$ for $r_{\text{c}} \in \{\hat{r}_{\text{c}}, r_{\text{c}}^\star\}$.

In practice, the calibrated proxy-reward model can be fit to a human-labeled preference dataset or to data annotated with calibrated true-rewards. Following (Huang et al., 2025), we assume for simplicity that $\hat{r}_{\text{c}}(y, x)$ is given.

Assumption 3.1 (Achievable maximum reward)**.** We assume that for $r_{\text{c}} \in \{\hat{r}_{\text{c}}, r_{\text{c}}^\star\}$, we have $r_{\text{c}}(\hat{y}(x), x) = 1$ for all $\hat{y}(x) \in \arg\max_y r_{\text{c}}(y, x)$ and given $x \in \mathcal{X}$.

In many settings, the calibrated reward function attains its maximum at specific responses. In particular, since a large language model (LLM) generates outputs using a finite vocabulary and a bounded number of tokens, the space of possible generations is finite, and thus the assumption holds trivially.

3.3 SBoN Algorithm

Fix a prompt $x \in \mathcal{X}$ and draw N i.i.d. candidates $Y_{1:N} \sim \pi_{\text{ref}}(\cdot \mid x)$. Let $Z \in \{1, \ldots, N\}$ denote the index of the selected response with distribution P_Z; write $P_Z(i) = \Pr(Z = i)$. We seek a distribution over indices that maximizes the calibrated proxy-reward:

$$\max_{P_Z \in \Delta_N} \; \mathbb{E}_Z\left[\hat{r}_{\text{c}}\big(Y_Z, x\big)\right].$$

Without regularization, the optimizer is the deterministic Best-of-N (BoN) rule $P_Z = \delta_{i^\star}$ with $i^\star \in \arg\max_i \hat{r}_{\text{c}}(Y_i, x)$. Because $\hat{r}_{\text{c}}$ is a *proxy* for the true-reward, this deterministic choice can overoptimize the proxy-reward and get response with lower calibrated true-reward. To smooth this, we add an entropy penalty with temperature $\beta > 0$:

$$\max_{P_Z \in \Delta_N} \; \mathbb{E}_Z\left[\hat{r}_{\text{c}}\big(Y_Z, x\big)\right] \; + \; \frac{1}{\beta} H(P_Z).$$

The unique solution is the softmax distribution, $P_Z(i) = \frac{\exp\big(\beta\, \hat{r}_{\text{c}}(Y_i, x)\big)}{\sum_{j=1}^N \exp\big(\beta\, \hat{r}_{\text{c}}(Y_j, x)\big)}$.

We then sample Z from this distribution and return Y_Z. We refer to this sampling rule as **Soft-BoN**, as introduced by Mayrink Verdun et al. (2025).

We denote the final policy from SBoN via $\pi_{\hat{r}_{\text{c}}}^{(N,\beta)}(y|x)$. Note that for $\beta \to \infty$ and $\beta \to -\infty$, we recover BoN and worst-of-N (WoN) (Balashankar et al., 2025), respectively. Furthermore, for $\beta \to 0$, we recover uniform sampling among the N response samples, which is equivalent to sampling from the reference model $\pi_{\text{ref}}(y|x)$. In (Mayrink Verdun et al., 2025, Lemma 1), the closed form solution of SBoN policy is derived,

$$\pi_{\hat{r}_{\text{c}}}^{(N,\beta)}(y|x) = \frac{\pi_{\text{ref}}(y|x)\exp(\beta \hat{r}_{\text{c}}(y, x))}{Z_{N,\beta}}, \tag{3}$$

where $Z_{N,\beta} = \mathbb{E}\Big[\Big(\frac{1}{N}\big(\exp(\beta \hat{r}_{\text{c}}(y, x)) + \sum_{i=1}^{N-1}\exp(\beta \hat{r}_{\text{c}}(Y_i, x))\big)\Big)^{-1}\Big]^{-1}$. Similarly, we can define $\pi_{r_{\text{c}}^\star}^{(N,\beta)}(y|x)$ based on a calibrated true-reward model. For simplicity, we define BoN policies under calibrated true-reward and proxy-reward models as $\pi_{r_{\text{c}}^\star}^{(N,\infty)}(y|x)$ and $\pi_{\hat{r}_{\text{c}}}^{(N,\infty)}(y|x)$, respectively. In this work, we focus on $\beta \geq 0$. Another motivation for SBoN based on the Gumbel-Max trick is provided in Appendix (App) D.

3.4 Tilted Error

Let's define the tilted error as the tilted average of square estimation error of calibrated true-reward function for a given prompt x with parameter β, as follows,

$$\varepsilon_{\beta, r_{\text{c}}}(x) := \frac{1}{\beta}\log\Big(\mathbb{E}_{Y \sim \pi_{\text{ref}}(y|x)}\big[e^{\beta(r_{\text{c}}^\star(Y,x) - \hat{r}_{\text{c}}(Y,x))^2}\big]\Big). \tag{4}$$

A similar definition of estimation error is introduced in (Yang & Wibisono, 2022). When $\beta = 0$, the definition reduces to the mean-squared error, which is also introduced in (Huang et al., 2025). Letting $\beta \to \infty$ recovers the square of the supremum (infinity) norm ($\|\cdot\|_\infty$) of the estimation error between $r^\star_{\mathrm{c}}(y,x)$ and $\hat{r}_{\mathrm{c}}(y,x)$. Therefore, the following properties hold for $\varepsilon_{\beta,r_{\mathrm{c}}}(x)$,

- The tilted error is bounded, i.e., $\varepsilon_{\beta,r_{\mathrm{c}}}(x) \in [0,1]$.
- The tilted average of the estimation error is monotonically increasing in β.
- $\varepsilon_{\infty,r}(x) := \lim_{\beta\to\infty} \varepsilon_{\beta,r}(x) = \|r^\star_{\mathrm{c}}(Y,x) - \hat{r}_{\mathrm{c}}(Y,x)\|^2_\infty$.

We assume that overoptimization regime happens whenever we have $\varepsilon_{\beta,r_{\mathrm{c}}}(x) > 0$.

We define tilted error using calibrated (proxy and true) reward models rather than raw reward models, because our focus is on how rankings change under the proxy. For example, if the proxy is a strictly increasing transform of the true-reward, the ranking is preserved; the Best-of-N (BoN) policy remains optimal and no overoptimization occurs. This behavior cannot be captured when working with the raw (uncalibrated) reward models. Note that in (Huang et al., 2025), raw (uncalibrated) reward models are utilized for error definition.

3.5 Optimal Policy and Coverage

We define the optimal policy under the calibrated true-reward model as,

$$\pi^\star_{r^\star_{\mathrm{c}}}(y|x) = \arg\max_{\pi} \mathbb{E}_{Y\sim\pi(\cdot|x)}[r^\star_{\mathrm{c}}(Y,x)]. \tag{5}$$

Similarly, we can define $\pi^\star_{\hat{r}}(y|x)$ as the optimal policy under the calibrated proxy-reward model.

As the calibrated reward functions (true and proxy) are bounded, we can interpret optimal policies as the limit of tilted optimal policies,

$$\pi_{\infty,r^\star_{\mathrm{c}}}(\cdot|x) := \lim_{\beta\to\infty} \pi_{\beta,r^\star_{\mathrm{c}}}(\cdot|x), \quad \pi_{\infty,\hat{r}_{\mathrm{c}}}(\cdot|x) := \lim_{\beta\to\infty} \pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x). \tag{6}$$

where $\pi_{\infty,r^\star_{\mathrm{c}}}(\cdot|x)$ and $\pi_{\infty,\hat{r}_{\mathrm{c}}}(\cdot|x)$ place all their probability mass on the maximizers of $r^\star_{\mathrm{c}}(y,x)$ and $\hat{r}_{\mathrm{c}}(y,x)$, respectively. Therefore we have $\pi^\star_{r^\star_{\mathrm{c}}}\big(\cdot\,|x\big) = \pi_{\infty,r^\star_{\mathrm{c}}}(\cdot|x)$ and $\pi^\star_{\hat{r}_{\mathrm{c}}}\big(\cdot\,|x\big) = \pi_{\infty,\hat{r}_{\mathrm{c}}}(\cdot|x)$.

Coverage: For a given calibrated reward function $r_{\mathrm{c}}(x,y)$, we define the tilted policy (softmax policy):

$$\pi_{\beta,r_{\mathrm{c}}}(y|x) \propto \pi_{\mathrm{ref}}(y|x)\exp(\beta r_{\mathrm{c}}(x,y)).$$

Then, we introduce the coverage of tilted policy with respect to the reference policy as,

$$C_{\beta,r_{\mathrm{c}},\mathrm{ref}}(x) := \sum_{y\in\mathcal{Y}} \frac{\pi^2_{\beta,r_{\mathrm{c}}}(y|x)}{\pi_{\mathrm{ref}}(y|x)}. \tag{7}$$

We also define,

$$C_{\infty,r_{\mathrm{c}},\mathrm{ref}}(x) := \lim_{\beta\to\infty} C_{\beta,r_{\mathrm{c}},\mathrm{ref}}(x).$$

This measure $C_{\beta,r_{\mathrm{c}},\mathrm{ref}}(x)$ can also be interpreted as a coverage constant, which is standard in KL-regularized policy learning. Furthermore, we can define the coverage of the tilted policy with respect to the reference policy as χ^2-divergence between $\pi_{\beta,r}(y|x)$ and $\pi_{\mathrm{ref}}(y|x)$, i.e., $\chi^2(\pi_{\beta,r}(y|x)\|\pi_{\mathrm{ref}}(y|x))$. It ensures that the reference policy places sufficient probability mass on high-reward responses, thereby guaranteeing that the support of the optimal policy lies within the support of the reference. This prevents cases where optimal outputs are entirely excluded by the reference. Similar notions of coverage have been explored in Huang et al. (2025).

3.6 Optimal Regret

For given policy $\pi\big(Y|x\big)$, we define expected calibrated true-reward with respect to the policy (a.k.a. value function[2]) as

$$J_{r^\star_{\mathrm{c}}}(\pi(\,\cdot\,|x)) := \mathbb{E}_{Y\sim\pi\left(\cdot|x\right)}[r^\star_{\mathrm{c}}(Y,x)]. \tag{8}$$

[2] We can also consider $\mathbb{E}_{X\sim\rho(\cdot)}[J_{r^\star_{\mathrm{c}}}(\pi(\cdot|X))]$. All of our results also holds for expected version of value function.

For two policies, $\pi_1(\cdot|x)$ and $\pi_2(\cdot|x)$, we define the gap between these two policies as follows,

$$\Delta_{J_{r_c^\star}}(\pi_1(\cdot|x), \pi_2(\cdot|x)) := J_{r_c^\star}(\pi_1(\cdot|x)) - J_{r_c^\star}(\pi_2(\cdot|x)). \tag{9}$$

We consider the following KL-Regularized objective function based on the calibrated true-reward function for SBoN,

$$J_{r_c^\star,\beta}(\pi_{\text{ref}}(y|x), \pi(\cdot|x)) := \mathbb{E}_{Y\sim\pi(\cdot|x)}[r_c^\star(Y,x)] - \frac{1}{\beta}\text{KL}(\pi(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)). \tag{10}$$

We provide an upper bound on the gap of the SBoN solution, which is the gap between $\pi^\star_{r_c^\star}(\,\cdot\,|x)$ as the optimal policy and $\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)$,

$$\Delta_{J_{r_c^\star}}(\pi^\star_{r_c^\star}(\,\cdot\,|x), \pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)) = J_{r_c^\star}(\pi^\star_{r_c^\star}(\,\cdot\,|x)) - J_{r_c^\star}(\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)). \tag{11}$$

Regarding regret of the BoN, we consider $\pi^{(N,\infty)}_{\hat{r}_c}(\cdot|x)$ instead of $\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)$ in equation 11.

Our results can be extended to the sub-optimal gap of the SBoN solution, which is the gap between $\pi_{\beta,r_c^\star}(\,\cdot\,|x)$ as the optimal solution to equation 10 and $\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)$,

$$\Delta_{J_{r_c^\star}}(\pi_{\beta,r_c^\star}(\,\cdot\,|x), \pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)) := J_{r_c^\star}(\pi_{\beta,r_c^\star}(\cdot|x)) - J_{r_c^\star}(\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)). \tag{12}$$

4 KL DIVERGENCE ANALYSIS

The KL divergence between the aligned policy and the reference policy, $\text{KL}(\pi^{(N,\infty)}_{r_c^\star}\|\pi_{\text{ref}})$, is studied by Beirami et al. (2024); Mroueh (2024) from a theoretical perspective. In particular, Beirami et al. (2024) derives an upper bound on KL divergence for BoN policies under the assumptions of a bijective true-reward mapping and a finite output space:

$$\text{KL}(\pi^{(N,\infty)}_{r_c^\star}(\cdot|x)\|\pi_{\text{ref}}(\cdot|x)) \le \log(N) - 1 + \frac{1}{N}, \tag{13}$$

Mroueh (2024) relaxes the bijectivity assumption and derives similar bounds using information-theoretic tools. Under some assumptions, the bound in equation 13 is tight. Furthermore, using Pinsker's inequality, in a similar approach to (Mroueh, 2024), we have,

$$\mathbb{E}_{Y\sim\pi^{(N,\beta)}_{r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] \le 0.5 + \sqrt{\frac{1}{2}\text{KL}(\pi^{(N,\beta)}_{r_c^\star}(\cdot|x)\|\pi_{\text{ref}}(\cdot|x))}. \tag{14}$$

Note that equation 14 implies that improvement of expected calibrated true-reward relative to the reference policy can not exceed the square root of the KL divergence. However, the analysis of KL divergence for the SBoN policy under the calibrated true-reward model is overlooked. Therefore, we first establish an upper bound on the KL divergence between the SBoN policy under the calibrated true-reward model and the reference policy, shedding light on its behaviour as a function of the number of samples N and temperature parameter β. All proof details are deferred to App. F.

Lemma 4.1. *The following upper bound holds on KL divergence between SBoN and reference policies for a given prompt $x \in \mathcal{X}$,*

$$\text{KL}(\pi^{(N,\beta)}_{r_c^\star}(y|x)\|\pi_{\text{ref}}(y|x)) \le \log\Big(\frac{N}{1+(N-1)\exp(-\beta)}\Big). \tag{15}$$

Using Lemma 4.1, we can observe that for BoN, $\beta \to \infty$, we have,

$$\text{KL}(\pi^{(N,\infty)}_{r_c^\star}(y|x)\|\pi_{\text{ref}}(y|x)) \le \log(N). \tag{16}$$

Comparing equation 16 with results in (Beirami et al., 2024; Mroueh, 2024), our result is derived from the SBoN asymptotic regime. Note that our bound is looser than the bound on KL divergence in equation 13. In contrast, our bound is general and can be applied to different β in SBoN. For $\beta = 0$, where our policy is the reference policy, our bound is tight. It is also important to note that the upper bound in Lemma 4.1 increases with the temperature parameter β for fixed N. Note that the result in Lemma 4.1 also holds for an arbitrary bounded reward model.

Recent works by Gao et al. (2023a) and Hilton et al. (2022) empirically demonstrate that, under a calibrated true-reward model, the improvement in expected calibrated true-reward, relative to a reference policy, scales proportionally to $\sqrt{\mathrm{KL}(\pi_{r_c^\star}^{(N,\infty)}\|\pi_{\mathrm{ref}})}$ for both RL and BoN policies. It is also observed by Gao et al. (2023b) that models optimized using proxy-rewards can suffer from overoptimization where the learned policy diverges further from the reference, the alignment may degrade. Despite theoretical advances, the KL divergence analysis for SBoN and BoN under the calibrated proxy-reward model remains largely unexplored. Therefore, we are interested in investigating the improvement of expected calibrated true-reward with respect to the SBoN policy under the calibrated proxy-reward model relative to the reference policy. For this purpose, we first propose the following useful Lemma to study the closeness of the SBoN policy under the calibrated true-reward model to the SBoN policy under the calibrated proxy-reward model in KL divergence measure.

Lemma 4.2. *The following upper bound holds on the KL divergence between the SBoN policies under calibrated true-reward and proxy-reward models respectively,*

$$\mathrm{KL}(\pi_{r_c^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}_c}^{(N,\beta)}(\cdot|x)) \leq \frac{N\beta\sqrt{\varepsilon_{\beta,r_c}(x)}}{1+(N-1)\exp(-\beta)}\Big(\frac{N\exp(2\beta)}{(N-1)^2}+1\Big). \tag{17}$$

Note that for $\beta = 0$, the upper bound in Lemma 4.2 is tight. Next, we derive our main result for improvement of the expected calibrated true-reward under the SBoN policy using the proxy model relative to the reference policy model.

Theorem 4.3. *The following upper bound holds on the improvement of expected calibrated true-reward under the SBoN policy for the calibrated proxy-reward model,*

$$\begin{aligned}\mathbb{E}_{Y\sim\pi_{\hat{r}_c}^{(N,\beta)}(\cdot|x)}[r_c^\star(Y,x)] \leq 0.5 + \sqrt{\frac{1}{2}\log\Big(\frac{N}{1+(N-1)\exp(-\beta)}\Big)} \\ + \min\Big(\sqrt{\frac{N\beta\sqrt{\varepsilon_{\beta,r_c}(x)}A(\beta,N)}{2(1+(N-1)\exp(-\beta))}},1\Big),\end{aligned} \tag{18}$$

where $A(\beta,N) = \Big(\frac{N\exp(2\beta)}{(N-1)^2}+1\Big)$.

Note that the results in Theorem 4.3 can also be interpreted as improvement relative to the reference policy as $\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[r_c^\star(Y,x)] = 0.5$. The upper bound in Theorem 4.3 includes two terms. The first term represents the upper bound on the expected calibrated true-reward under the SBoN policy relative to the reference policy; we are interested in maximizing this term. Note that, as mentioned in (Gao et al., 2023b), the expected calibrated true-reward under the aligned policy, relative to the reference policy, is proportional to the square root of KL divergence. The second term quantifies the estimation error introduced by substituting a calibrated proxy-reward model for the calibrated true-reward model. We aim to minimize the second term.

Next, we compare BoN and SBoN under the no-reward model and reward model overoptimization.

Remark 4.4 (No overoptimization). We can observe that for a given β, if we assume $\varepsilon_{\beta,r_c}(x) = 0$, then we have $\mathrm{KL}(\pi_{r_c^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}_c}^{(N,\beta)}(\cdot|x)) = 0$. Under this assumption, the upper bound in Theorem 4.3 simplifies to $\sqrt{\frac{1}{2}\log\Big(\frac{N}{1+(N-1)\exp(-\beta)}\Big)}$ which is monotonically increasing in β. Because a larger KL divergence is desirable in this context, as proposed by (Gao et al., 2023b), the BoN policy is preferred under no overoptimization scenario.

Remark 4.5 (Overoptimization). When $\varepsilon_{\beta,r_c}(x) > 0$, we have two conflicting goals in Theorem 4.3: one suggesting for fixed N that β needs to be smaller for better estimation of the true policy by the calibrated proxy-reward model one given in second term of equation 18, and another one suggesting a larger β to induce a better upper bound on the expected calibrated true-reward in first term of equation 18. Hence, for a given N, there exists an optimal β to balance between the estimation error term and the expected calibrated true-reward under the SBoN policy for the calibrated true-reward model. In this scenario, SBoN can lead to better tradeoffs than BoN. A similar discussion can be done for fixed β and varying N.

5 Regret Analysis

In this section, we derive theoretical regret bounds for SBoN and BoN based on calibrated reward models. First, we provide a helpful Lemma regarding the expected coverage assumption that can help us interpret the results of regret for BoN and SBoN. All proof details are deferred to App. G.

Lemma 5.1. *Under Assumption 3.1, it holds that* $C_{\infty,r_{\mathrm{c}},\mathrm{ref}}(x) = \frac{1}{\sum_i \pi_{\mathrm{ref}}(y_{i,r}^{\max}(x)|x)}$, *where* $y_{i,r}^{\max}(x) \in \arg\max_y r_{\mathrm{c}}(y,x)$.

Now, we derive an upper bound on the regret of SBoN.

Theorem 5.2 (Optimal Gap of SBoN)**.** *Under Assumption 3.1, the following upper bound holds on the optimal regret gap of the SBoN policy for any* $\beta > 0$,

$$\Delta_{J_{r_{\mathrm{c}}^\star}}(\pi^\star_{r_{\mathrm{c}}^\star}(\,\cdot\,|x), \pi_{\hat{r}_{\mathrm{c}}}^{(N,\beta)}(\cdot|x)) \le \sqrt{\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\Big(\sqrt{C_{\infty,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)}\Big) + 2\sqrt{\frac{1}{2}\log\left(1+\frac{C_{\infty,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)-1}{N}\right)} + \frac{\log(C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x))}{\beta},$$

Regret of BoN Through Smoothing Lens: We now derive an upper bound on the regret of BoN by taking the asymptotic limit of the regret bound on optimal gap of SBoN in Theorem 5.3.

Theorem 5.3 (Optimal Gap of BoN)**.** *Under Assumption 3.1, the following upper bound holds on the optimal regret gap of the BoN policy for any* $\beta > 0$,

$$\Delta_{J_{r_{\mathrm{c}}^\star}}(\pi^\star_{r_{\mathrm{c}}^\star}(\cdot|x), \pi_{\hat{r}_{\mathrm{c}}}^{(N,\infty)}(\cdot|x)) \le \sqrt{\varepsilon_{\infty,r}(x)}\Big(\sqrt{C_{\infty,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)}\Big) + 2\sqrt{\frac{1}{2}\log\left(1+\frac{C_{\infty,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)-1}{N}\right)}.$$

Remark 5.4 (Comparison with (Huang et al., 2025)). The regret bound for BoN policy grows with the L_∞-norm of the reward-model estimation error. In contrast to the result in (Huang et al., 2025), our bound remains finite whenever the overoptimization error vanishes, i.e., when $\varepsilon_{\infty,\beta}(x) = 0$ or N grows. We also derive results based on calibrated reward, instead of raw (uncalibrated) reward models.

Remark 5.5 (Quality of reference policy). Furthermore, the bound stated in Theorem 5.3 (or Theorem 5.2) depends on the quantity, $C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x) = \frac{1}{\sum_i \pi_{\mathrm{ref}}\left(y_{i,r^\star}^{\max}(x)|x\right)}$, where $y_{i,r^\star}^{\max}(x) \in \arg\max_y r_{\mathrm{c}}^\star(y,x)$. It can be interpreted as the quality of reference policy under the calibrated true-reward model. Similarly, the bound in Theorem 5.3 (or Theorem 5.2) depends on the quantity, $C_{\infty,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x) = \frac{1}{\sum_i \pi_{\mathrm{ref}}\left(y_{i,\hat{r}}^{\max}(x)|x\right)}$, where $y_{i,\hat{r}}^{\max}(x) \in \arg\max_y \hat{r}_{\mathrm{c}}(y,x)$. It can be interpreted as quality of reference policy under the calibrated proxy-reward model. Therefore, the quality of reference policy under both calibrated true-reward and proxy-reward models affect the performance of BoN and SBoN policies.

Next, we compare how BoN and SBoN perform when overoptimization is present and when it is absent.

Remark 5.6 (Overoptimization). Assume that the calibrated proxy-reward suffers from overoptimization, i.e. $\varepsilon_{\beta,r_{\mathrm{c}}}(x) > 0$ for every $\beta > 0$. Letting $N \to \infty$ and invoking Theorem 5.2, we obtain

$$\Delta_{J_{r_{\mathrm{c}}^\star}}\big(\pi^\star_{r_\star}(\cdot\mid x), \pi_{\hat{r}_{\mathrm{c}}}^{(\infty,\beta)}(\cdot\mid x)\big) \le \frac{\log C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)}{\beta} + \sqrt{\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\Big(\sqrt{C_{\infty,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)}\Big). \tag{19}$$

Similarly, for BoN we have,

$$\Delta_{J_{r_{\mathrm{c}}^\star}}\big(\pi^\star_{r_\star}(\cdot\mid x), \pi_{\hat{r}_{\mathrm{c}}}^{(\infty,\infty)}(\cdot\mid x)\big) \le \sqrt{\varepsilon_{\infty,r}(x)}\Big(\sqrt{C_{\infty,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)} + \sqrt{C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)}\Big). \tag{20}$$

Define the auxiliary function

$$g(\beta) \;=\; \beta\big(\varepsilon_{\infty,r}(x) - \varepsilon_{\beta,r_{\mathrm{c}}}(x)\big), \qquad \beta \ge 0.$$

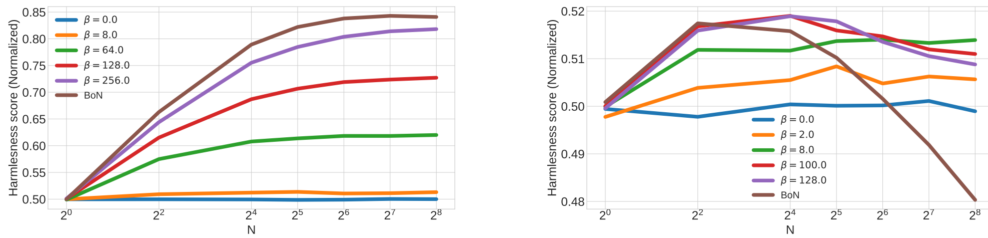

Figure 1: Soft Best-of-N experiment using a strong reward model (Left) and a weak one (Right). When the performance of the reward model is poor, BoN can lead to overoptimization, while the SBoN can help to mitigate it.

Because $g(0) = g(\infty) = 0$ and $g(\beta) \geq 0$ for all β, there exists at least one maximizer $\beta^\star \in (0, \infty)$ such that $g(\beta^\star) = \max_{\beta \geq 0} g(\beta)$.

If $\frac{\log C_{\infty, r_c^\star, \mathrm{ref}}(x)}{\sqrt{C_{\beta, \hat{r}_c, \mathrm{ref}}(x)} + \sqrt{C_{\infty, r_c^\star, \mathrm{ref}}(x)}} \leq g(\beta^\star)$, then the upper bound in equation 19 does not exceed equation 20, and hence the bound on the regret of the SBoN policy is *tighter* than the bound on the regret of the BoN policy under the calibrated proxy-reward model. An analogous comparison can be carried out for any fixed β and changing N.

Remark 5.7 (No overoptimization). Assume that the overoptimization vanishes, i.e. $\varepsilon_{\beta, r_c}(x) = 0$ for every $\beta \in [0, \infty)$. Then the optimality gaps of the SBoN and BoN policies satisfy

$$\Delta_{J_{r_c^\star}}\big(\pi^\star_{r_\star}(\cdot \mid x), \pi^{(N,\beta)}_{\hat{r}_c}(\cdot \mid x)\big) \leq 2\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty, \hat{r}_c, \mathrm{ref}}(x) - 1}{N}\Big)} + \frac{\log C_{\infty, r_c^\star, \mathrm{ref}}(x)}{\beta}, \tag{21}$$

$$\Delta_{J_{r_c^\star}}\big(\pi^\star_{r_\star}(\cdot \mid x), \pi^{(N,\infty)}_{\hat{r}_c}(\cdot \mid x)\big) \leq 2\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty, \hat{r}_c, \mathrm{ref}}(x) - 1}{N}\Big)}. \tag{22}$$

By Lemma 5.1, $C_{\infty, r_c^\star, \mathrm{ref}}(x) \geq 1$; consequently, the bound in equation 22 is tighter than the bound in equation 21.

6 Empirical Evidence

To support our theoretical analysis, we conducted experiments comparing Soft Best-of-N (SBoN) across different regularization strengths and reward model qualities. We used the Olmo-2 1B model (OLMo et al., 2024) as the generator and prompts from the Attaq dataset (Kour et al., 2023). For each prompt, we generated multiple responses and selected one using SBoN with varying temperature values β. We ran two experimental conditions: one using a strong proxy-reward model (ArmoRM 8B (Wang et al., 2024)) which is close to true-reward model, and another using a weaker proxy-reward model (Beaver 7B RM (Dai et al., 2023)). We use LLM-as-a-Judge Zheng et al. (2023) as our r^*. As shown in Figure 1, when the reward model is weak, performance degrades for large N due to reward hacking. However, the smoothing in SBoN helps mitigate this degradation. This observation is also aligned with our theoretical analysis and discussion in Section 4, where under overoptimization there exists a β for a given N which outperforms BoN. For more details, see App. H. We also studied the behavior of our upper bound on the KL divergence between the SBoN policy and the reference policy, Lemma 4.2, at App. H.2. More experiments with another reward model are provided in App. H.1.

7 Conclusion

In this work, we establish a theoretical foundation for alignment strategies based on Soft Best-of-N (SBoN) and Best-of-N (BoN) policies. Specifically, we derive upper bounds on the KL divergence between the aligned policysuch as SBoN or BoNand the reference policy. We also studied the regret gap between the optimal policy and the aligned policy, e.g., BoN and SBoN policies. We further analyze how errors in reward estimation affect performance in both KL divergence and regret gap. Notably, both our theoretical analysis and empirical results demonstrate that, under a calibrated proxy-reward model where overoptimization happens, SBoN perform better than BoN under some conditions.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Afra Amini, Tim Vieira, Elliott Ash, and Ryan Cotterell. Variational best-of-n alignment. In *International Conference on Learning Representations (ICLR)*, 2025.

Gholamali Aminian, Amir R Asadi, Idan Shenfeld, and Youssef Mroueh. Theoretical analysis of kl-regularized rlhf with multiple reference models. *arXiv preprint arXiv:2502.01203*, 2025.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

Ananth Balashankar, Ziteng Sun, Jonathan Berant, Jacob Eisenstein, Michael Collins, Adrian Hutter, Jong Lee, Chirag Nagpal, Flavien Prost, Aradhana Sinha, Ananda Theertha Suresh, and Ahmad Beirami. Infalign: Inference-aware language model alignment. *International Conference on Machine Learning (ICML)*, 2025.

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. *International Conference on Machine Learning (ICML)*, 2024.

Vivek S Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 2002.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Clément L Canonne. A short note on an inequality between kl and tv. *arXiv preprint arXiv:2202.07198*, 2022.

Pierre Cardaliaguet, François Delarue, Jean-Michel Lasry, and Pierre-Louis Lions. *The master equation and the convergence problem in mean field games:(ams-201)*. Princeton University Press, 2019.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

L Gao, J Tow, B Abbasi, S Biderman, S Black, A DiPofi, C Foster, L Golding, J Hsu, A Le Noach, et al. A framework for few-shot language model evaluation. *URL https://zenodo.org/records/10256836*, 7, 2023a.

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

Jonathan Geuter, Youssef Mroueh, and David Alvarez-Melis. Guided speculative inference for efficient test-time alignment of llms. *arXiv preprint arXiv:2506.04118*, 2025.

Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832*, 2024.

Emil Julius Gumbel. Statistical theory of extreme values and some practical applications. *Nat. Bur. Standards Appl. Math. Ser. 33*, 1954.

J. Hilton, P. Clark, et al. Measuring goodharts law: Towards an evaluation framework for open-ended generative models. https://openai.com/index/measuring-goodharts-law, 2022. Accessed: 2025-01-30.

Ronald A Howard and James E Matheson. Risk-sensitive markov decision processes. *Management science*, 1972.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Dylan J Foster, and Akshay Krishnamurthy. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*, 2024.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Eiji Uchibe. Evaluation of best-of-n sampling strategies for language model alignment. *Transactions on Machine Learning Research*.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hanna Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633, 2024.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling to mitigate reward hacking for language model alignment. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, 2024.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio du Pin Calmon. Inference-time reward hacking in large language models. *arXiv preprint arXiv:2506.19248*, 2025.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. *arXiv preprint arXiv:2402.01694*, 2024.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. *Advances in Neural Information Processing Systems*, 35:16203–16220, 2022a.

Tomasz Korbak, Ethan Perez, and Christopher L Buckley. Rl with kl penalties is better viewed as bayesian inference. *arXiv preprint arXiv:2205.11275*, 2022b.

Barry W Kort and Dimitri P Bertsekas. A new penalty function method for constrained minimization. In *IEEE Conference on Decision and Control and 11th Symposium on Adaptive Processes*, 1972.

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.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *Journal of Machine Learning Research*, 24(142):1–79, 2023.

Guan-Horng Liu and Evangelos A Theodorou. Deep learning theory review: An optimal control and dynamical systems perspective. *arXiv preprint arXiv:1908.10920*, 2019.

Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio P Calmon. Soft best-of-n sampling for model alignment. *arXiv preprint arXiv:2505.03156*, 2025.

Youssef Mroueh. Information theoretic guarantees for policy alignment in large language models. *arXiv preprint arXiv:2406.05883*, 2024.

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. In *International Conference on Machine Learning*, pp. 36486–36503. PMLR, 2024.

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.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

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.

EY Pee and Johannes O Royset. On solving large-scale finite minimax problems using exponential smoothing. *Journal of Optimization Theory and Applications*, 2011.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022.

Jiahao Qiu, Yifu Lu, Yifan Zeng, Jiacheng Guo, Jiayi Geng, Huazheng Wang, Kaixuan Huang, Yue Wu, and Mengdi Wang. Treebon: Enhancing inference-time alignment with speculative tree-search and best-of-n sampling. *arXiv preprint arXiv:2410.16033*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, et al. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622*, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yuda Song, Gokul Swamy, Aarti Singh, Drew Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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.

Benedikt Stroebl, Sayash Kapoor, and Arvind Narayanan. Inference scaling f-laws: The limits of llm resampling with imperfect verifiers. *arXiv preprint arXiv:2411.17501*, 2024.

Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. *arXiv preprint arXiv:2410.20290*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP*, 2024.

Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 2013.

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. In *Forty-first International Conference on Machine Learning*, 2024.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. Asymptotics of language model alignment. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 2027–2032. IEEE, 2024.

Kaylee Yingxi Yang and Andre Wibisono. Convergence of the inexact langevin algorithm and score-based generative models in kl divergence. *arXiv preprint arXiv:2211.01512*, 2022.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*, 2024.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

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*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

A Other Related Works

Smoothing of Maximum: Approximating the maximum operator using a smoothed or softmax-based surrogate is a widely adopted technique in machine learning. This approach is particularly useful in settings where the hard maximum is non-differentiable or leads to unstable optimization. For instance, in robust regression, smooth approximations to the max operator are used in min-max formulations to achieve tractable optimization under distributional shifts (Wang et al., 2013; Li et al., 2023). In sequential decision-making, similar ideas appear in risk-sensitive control and Q-learning, where the softmax of Q-values leads to stochastic policies that balance exploration and exploitation (Howard & Matheson, 1972; Borkar, 2002). In convex and non-convex optimization, smoothing the maximum objective has been shown to improve convergence properties (Kort & Bertsekas, 1972; Pee & Royset, 2011; Liu & Theodorou, 2019). The Soft Best-of-N (SBoN) framework, (Mayrink Verdun et al., 2025; Khanov et al., 2024; Jinnai et al., 2024), leverages this principle by replacing the hard selection of the highest-reward sample with a softmax-weighted sampling distribution. Regarding the SBoN, the empirical version of SBoN is introduced by (Khanov et al., 2024) as ARGS-stochastic, where a token from a probability distribution among the top-k candidate tokens is chosen. Then, the regularized version of BoN, which can be represented as SBoN, is discussed by (Jinnai et al., 2024). Given the broad success of SBoN, we are motivated to theoretically investigate the SBoN policies and the effect of the calibrated proxy-reward model (reward hacking) and the quality of the reference policy.

Theoretical Foundation of RLHF: Several works have studied the theoretical underpinnings of reverse KL-regularized RLHF, particularly in terms of sample complexity (Zhao et al., 2024; Xiong et al., 2024; Song et al., 2024; Zhan et al., 2023; Ye et al., 2024; Aminian et al., 2025). Note that, as the sampling distributions in BoN and SBoN are different, we can not apply RLHF analysis to these sampling strategies. Therefore, it is needed to develop new foundations for BoN and SBoN.

Overoptimization. Alignment methods are widely known to suffer from overoptimization, also known as misspecification, reward hacking, or Goodhart Law, where optimizing against a calibrated proxy-reward model leads to degraded performance compared to the calibrated true-reward model (Amodei et al., 2016; Casper et al., 2023; Gao et al., 2023b). This issue is particularly pronounced in inference-time alignment methods such as BoN, where an increasing number of responses N makes the overoptimization problem worse (Huang et al., 2025; Stroebl et al., 2024; Gao et al., 2023b). Huang et al. (2025) theoretically demonstrate that the BoN policy suffers from overoptimization when N is large, given a fixed estimation error in the reward model, and propose a solution based on a χ^2-regularized framework. Other approaches to mitigating this issue include ensembling strategies (Coste et al.; Eisenstein et al.) and regularization techniques (Ichihara et al.). In a concurrent line of work, Khalaf et al. (2025) introduce the Best-of-Poisson method to reduce overoptimization in inference-time algorithms. The overoptimization in BoN and SBoN is also studied by Khalaf et al. (2025) and a principled hedging framework is proposed to mitigate the overoptimization. In contrast, we study overoptimization in inference-time alignment methods SBoN and BoN from the perspectives of regret gap and KL divergence analysis.

B Calibrated Reward

Inspired by (Balashankar et al., 2025), in this section, we provide more details regarding calibrated reward. A standard metric for evaluation of models is the *win-rate* relative to a base policy π_{ref} (Stiennon et al., 2020; Gao et al., 2023b). For a prompt x and responses y, z, define the win random variable under raw (uncalibrated) reward r as

$$w_r(y, z \mid x) = \mathbf{1}\{r(y, x) > r(z, x)\} + \tfrac{1}{2}\,\mathbf{1}\{r(y, x) = r(z, x)\}.$$

Definition B.1 (Calibrated reward)**.** The calibrated reward of y under policy π is its expected win-rate probability against $z \sim \pi(\cdot \mid x)$:

$$r_{\mathrm{c},\pi}(x, y) \coloneqq \mathbb{E}_{z\sim\pi(\cdot|x)}[w_r(y, z \mid x)]\,.$$

In practice, we consider $\pi = \pi_{\text{ref}}$, therefore we denote calibrated reward via $r_{\mathrm{c}}(x, y)$ under reference policy. In the following, we provide some reasons for choosing calibrated reward instead of raw (uncalibrated) reward in our work,

- **Matches win-rate evaluation.** For any policies π_1, π_2,
$$\mathrm{W}_r(\pi_1 \succ \pi_2 \mid x) := \mathbb{E}_{y\sim\pi_1(\cdot|x)}[\, r_{c,\pi_2}(x,y)] \,,$$
where $\mathrm{W}_r(\pi_1 \succ \pi_2 \mid x)$ is standard win-rate. So maximizing $\mathbb{E}_{y\sim\pi}[r_{c,\pi_{\mathrm{ref}}}(x,y)]$ directly optimizes standard win rate vs. the base model.
- **Invariance to score scaling.** If m is strictly increasing and $r' = m \circ r$, then
$$r'_c(x,y) \;=\; r_c(x,y),$$
making the target robust to arbitrary monotone reparameterizations of the reward (e.g., affine rescaling, temperature).
- **Unified, probabilistic scale.** For $y \sim \pi_{\mathrm{ref}}(\cdot \mid x)$,
$$r_c(x,y) \sim \mathrm{Unif}[0,1],$$
independent of both r and π_{ref}. This normalizes per-prompt reward scales and interprets scores as win probabilities.

C Summary of KL divergence Results

The connections between optimal, SBoN, BoN and tilted optimal policies under true or proxy-reward models are shown in Figure 2.

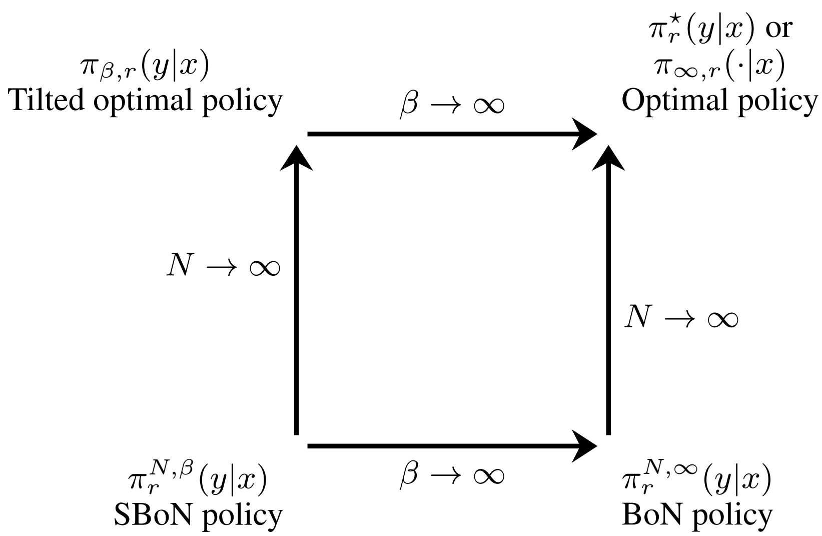

Figure 2: Connections of different policies under reward model $r \in \{\hat{r}_{\mathrm{c}}(y,x), r_{\mathrm{c}}^\star(y,x)\}$

In Table 1 , we summarize results on KL divergences between the aligned and reference policies, along with corresponding upper bounds for both SBoN and BoN policies. Furthermore, in Table 2, we summarize results on KL divergences between aligned policies under true and proxy-reward models, along with upper bounds for SBoN and tilted policies.

D GumbelMax trick

We also provide an interpretation for SBoN from the Gumbel-Max trick. An alternative way to sample Z from
$$\Pr(Z=i) \;\propto\; \exp\big(\beta\, \hat{r}_{\mathrm{c}}(Y_i,x)\big)$$
is via the GumbelMax trick. We can draw independent Gumbel-distributed random variables $G_i \sim \mathrm{Gumbel}(0,1), i = 1,\ldots,n$, and then set
$$Z = \arg\max_{i\in\{1,\ldots,N\}} \left[\hat{r}_{\mathrm{c}}(Y_i,x) + \frac{G_i}{\beta}\right].$$

Table 1: KL divergences between the aligned and reference policies, along with corresponding upper bounds for both SBoN and BoN policies.

KL divergence Term	Theorem / Lemma	Upper Bound
$\mathrm{KL}\big(\pi_{r_c^\star}^{(N,\beta)}(\cdot\|x)\,\|\,\pi_{\mathrm{ref}}(\cdot\|x)\big)$	Lemma 4.1	$\log\left(\frac{N}{1+(N-1)\exp(-\beta)}\right)$
$\mathrm{KL}(\pi_{r_c^\star}^{(N,\infty)}(\cdot\|x)\|\pi_{\mathrm{ref}}(\cdot\|x))$	Theorem 3.1 in (Beirami et al., 2024) and Theorem 1 in (Mroueh, 2024)	$\log(N)-1+1/N$

Table 2: KL divergences between aligned policies under true and proxy-reward models, along with upper bounds for SBoN and tilted policies.

KL divergence Term	Theorem / Lemma	Upper Bound
$\mathrm{KL}\big(\pi_{r_c^\star}^{(N,\beta)}(\cdot\|x)\,\|\,\pi_{\hat{r}_c}^{(N,\beta)}(\cdot\|x)\big)$	Lemma 4.2	$\frac{N\beta\sqrt{\varepsilon_{\beta,r_c}(x)}}{1+(N-1)\exp(-\beta)}\left(\frac{N\exp(2\beta)}{(N-1)^2}+1\right)$
$\mathrm{KL}\big(\pi_{\beta,r_c^\star}(\cdot\|x)\,\|\,\pi_{\beta,\hat{r}_c}(\cdot\|x)\big)$	Lemma E.8	$2\beta\sqrt{\varepsilon_{\beta,r_c}(x)}\left(\sqrt{\frac{\mathbb{E}[\exp(2\beta\hat{r}_c(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}_c(Y,x))]}}+\sqrt{\frac{\mathbb{E}[\exp(2\beta r_c^\star(Y,x))]}{\mathbb{E}^2[\exp(\beta r_c^\star(Y,x))]}}\right)$

By properties of the Gumbel distribution, this yields exactly the same softmax sampling law, without needing to compute the normalizing factor $\sum_{j=1}^{N}\exp(\beta\hat{r}(Y_j,x))$ explicitly (Gumbel, 1954). When $\beta\to\infty$, the effect of the Gumbel noises vanishes and the sampling strategy reduces to BoN.

E Technical Tools

We denote the set maximizers of the calibrated proxy-reward function via $\hat{\mathcal{Y}}(x)=\{\hat{y}_j(x)\}_{j=1}^{m(x)}$.

We introduce the functional derivative, see Cardaliaguet et al. (2019).

Definition E.1. (Cardaliaguet et al., 2019) A functional $U:\mathcal{P}(\mathbb{R}^n)\to\mathbb{R}$ admits a functional derivative if there is a map $\frac{\delta U}{\delta m}:\mathcal{P}(\mathbb{R}^n)\times\mathbb{R}^n\to\mathbb{R}$ which is continuous on $\mathcal{P}(\mathbb{R}^n)$ and, for all $m,m'\in\mathcal{P}(\mathbb{R}^n)$, it holds that

$$U(m')-U(m)=\int_0^1\int_{\mathbb{R}^n}\frac{\delta U}{\delta m}(m_\beta,a)\,(m'-m)(da)\,\mathrm{d}\beta,$$

where $m_\beta=m+\beta(m'-m)$.

Definition E.2 (Sensitivity of a policy)**.** We also define the sensitivity of a policy $\pi_r(y|x)$, which is a function of reward function $r(x,y)$, with respect to the reward function as

$$\frac{\partial\pi}{\partial r}(r):=\lim_{\Delta r\to 0}\frac{\pi_r(y|x)-\pi_{r+\Delta r}(y|x)}{\Delta r}. \tag{23}$$

Lemma E.3 (Kantorovich-Rubenstein duality of total variation distance, see (Polyanskiy & Wu, 2022))**.** *The Kantorovich-Rubenstein duality (variational representation) of the total variation distance is as follows:*

$$\mathbb{TV}(m_1,m_2)=\frac{1}{2L}\sup_{g\in\mathcal{G}_L}\left\{\mathbb{E}_{Z\sim m_1}[g(Z)]-\mathbb{E}_{Z\sim m_2}[g(Z)]\right\}, \tag{24}$$

where $\mathcal{G}_L=\{g:\mathcal{Z}\to\mathbb{R},||g||_\infty\le L\}$.

Lemma E.4 (Lemma 5.4 in (Aminian et al., 2025))**.** *Consider the softmax policy, $\pi_r^\beta(y|x)\propto\pi_{\mathrm{ref}}(y|x)\exp(\beta r(x,y))$. Then, the sensitivity of the policy with respect to the reward function is*

$$\frac{\partial\pi_r^\beta}{\partial r}(r)=\beta\pi_r^\beta(y|x)(1-\pi_r^\beta(y|x)).$$

Lemma E.5 (Pinskers Inequality (Canonne, 2022))**.** *For m_1 and m_2, we have,*

$$\mathbb{TV}(m_1,m_2)\le\sqrt{\frac{1}{2}\mathrm{KL}(m_2\|m_1)}\,. \tag{25}$$

The following Lemmata are useful for our technical proofs.

Lemma E.6. *The following upper bound holds,*

$$\log\left(\frac{Z_{r^\star,Y}(x,\beta)}{Z_{\hat{r},Y}(x,\beta)}\right) \leq \beta\sqrt{\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\sqrt{C_{\beta,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)}. \tag{26}$$

Proof.

$$\begin{aligned}
\frac{Z_{\hat{r}_{\mathrm{c}},Y}(x,\beta)}{Z_{r_{\mathrm{c}}^\star,Y}(x,\beta)} &= \frac{\sum_{\mathcal{Y}}\exp(\beta\hat{r}_{\mathrm{c}}(y,x))\pi_{\mathrm{ref}}(y|x)}{\sum_{\mathcal{Y}}\exp(\beta r_{\mathrm{c}}^\star(y,x))\pi_{\mathrm{ref}}(y|x)} \\
&= \frac{\sum_{\mathcal{Y}}\exp(\beta(\hat{r}_{\mathrm{c}}(y,x)-r_{\mathrm{c}}^\star(y,x)))\exp(\beta r_{\mathrm{c}}^\star(y,x))\pi_{\mathrm{ref}}(y|x)}{\sum_{\mathcal{Y}}\exp(\beta r_{\mathrm{c}}^\star(y,x))\pi_{\mathrm{ref}}(y|x)} \\
&= \sum_{\mathcal{Y}}\pi_{\beta,r_{\mathrm{c}}^\star}(y|x)\exp(\beta(\hat{r}_{\mathrm{c}}(y,x)-r_{\mathrm{c}}^\star(y,x)))
\end{aligned} \tag{27}$$

Due to convexity of $-\log(\cdot)$ and using CauchySchwarz inequality, we have,

$$\begin{aligned}
-\log\left(\frac{Z_{\hat{r}_{\mathrm{c}},Y}(x,\beta)}{Z_{r_{\mathrm{c}}^\star,Y}(x,\beta)}\right) &\leq \beta\sum_{\mathcal{Y}}\pi_{\beta,r_{\mathrm{c}}^\star}(y|x)(r_{\mathrm{c}}^\star(y,x)-\hat{r}_{\mathrm{c}}(y,x)) \\
&\leq \beta\sum_{\mathcal{Y}}\frac{\pi_{\beta,r_{\mathrm{c}}^\star}(y|x)}{\pi_{\mathrm{ref}}(y|x)}(r_{\mathrm{c}}^\star(y,x)-\hat{r}_{\mathrm{c}}(y,x))\pi_{\mathrm{ref}}(y|x) \\
&\leq \beta\sqrt{\sum_{\mathcal{Y}}(r_{\mathrm{c}}^\star(y,x)-\hat{r}_{\mathrm{c}}(y,x))^2\pi_{\mathrm{ref}}(y|x)}\sqrt{\sum_{\mathcal{Y}}\frac{\pi_{\beta,r_{\mathrm{c}}^\star}^2(y|x)}{\pi_{\mathrm{ref}}(y|x)}} \\
&= \sqrt{\beta}\sqrt{\sum_{\mathcal{Y}}\log\big(\exp\big(\beta(r_{\mathrm{c}}^\star(y,x)-\hat{r}_{\mathrm{c}}(y,x))^2\big)\big)\pi_{\mathrm{ref}}(y|x)}\sqrt{C_{\beta,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)} \\
&\leq \beta\sqrt{\frac{1}{\beta}\log\big(\sum_{\mathcal{Y}}\exp\big(\beta(r_{\mathrm{c}}^\star(y,x)-\hat{r}_{\mathrm{c}}(y,x))^2\big)\pi_{\mathrm{ref}}(y|x)\big)}\sqrt{C_{\beta,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)} \\
&= \beta\sqrt{\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\sqrt{C_{\beta,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)},
\end{aligned} \tag{28}$$

□

Lemma E.7. *The following holds,*

$$\mathrm{KL}(\pi_{r_{\mathrm{c}}^\star}^\star(\,\cdot\,|x)\|\pi_{\mathrm{ref}}(\cdot|x)) \leq \log(C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)) \tag{29}$$

Proof. Note that, we have,

$$\begin{aligned}
\mathrm{KL}(\pi_{r_{\mathrm{c}}^\star}^\star(\,\cdot\,|x)\|\pi_{\mathrm{ref}}(\cdot|x)) &\leq \log\left(\mathbb{E}_{Y\sim\pi_{r_{\mathrm{c}}^\star}^\star(\cdot|x)}\left[\frac{\pi_{r_{\mathrm{c}}^\star}^\star(\,\cdot\,|x)}{\pi_{\mathrm{ref}}(\cdot|x)}\right]\right) \\
&\leq \log(C_{\infty,r_{\mathrm{c}}^\star,\mathrm{ref}}(x))
\end{aligned} \tag{30}$$

□

Lemma E.8. *The following upper bound holds,*

$$\mathrm{KL}(\pi_{\beta,r_{\mathrm{c}}^\star}(y|x)\|\pi_{\beta,\hat{r}_{\mathrm{c}}}(y|x)) \leq \beta\sqrt{\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\big(\sqrt{C_{\beta,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)}+\sqrt{C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)}\big). \tag{31}$$

Proof.

$$\begin{aligned}
\mathrm{KL}(\pi_{\beta,r_c^\star}(y|x)\|\pi_{\beta,\hat{r}_c}(y|x)) &= \sum_{\mathcal{Y}} \pi_{\beta,r_c^\star}(y|x)\log\Big(\frac{\pi_{\beta,r_c^\star}(y|x)}{\pi_{\beta,\hat{r}_c}(y|x)}\Big) \\
&= \beta\sum_{\mathcal{Y}}(r_c^\star(y,x)-\hat{r}_c(y,x))\pi_{\beta,r_c^\star}(y|x) + \log(Z_{\hat{r},Y}(x,\beta)/Z_{r^\star,Y}(x,\beta)) \\
&\le \beta\sqrt{\varepsilon_{\beta,r_c}(x)}\big(\sqrt{C_{\beta,r_c^\star,\mathrm{ref}}(x)}+\sqrt{C_{\beta,\hat{r}_c,\mathrm{ref}}(x)}\big),
\end{aligned} \tag{32}$$

where the final inequality holds due to Lemma E.6 and applying CauchySchwarz inequality. □

Lemma E.9. *Suppose that $f(Z)\in[0,Z_{\max}]$, $\mathcal{Z}_{\max}=\{z_{m,i}\}_{i=1}^m$ is the set of miximizers of $f(Z)$, i.e., $f(z)=Z_{\max}$ for $z\in\mathcal{Z}_{\max}$. Then we have,*

$$\lim_{\beta\to\infty}\frac{\mathbb{E}[\exp(2\beta f(Z))]}{\mathbb{E}[\exp(\beta f(Z))]^2} = \frac{1}{\sum_{z\in\mathcal{Z}_{\max}}P(Z=z)}. \tag{33}$$

Proof.

$$\frac{\mathbb{E}[\exp(2\beta f(Z))]}{\mathbb{E}[\exp(\beta f(Z))]^2} = \frac{\mathbb{E}[\exp(2\beta(f(Z)-Z_{\max}))]}{\mathbb{E}[\exp(\beta(f(Z)-Z_{\max}))]^2} \tag{34}$$

$$\frac{\sum_j P(Z=z_j)\exp(2\beta(f(z_j)-Z_{\max}))}{(\sum_j P(Z=z_j)\exp(\beta(f(z_j)-Z_{\max})))^2} \tag{35}$$

Now, we have,

$$\begin{aligned}
&\lim_{\beta\to\infty}\frac{\mathbb{E}[\exp(2\beta f(Z))]}{\mathbb{E}[\exp(\beta f(Z))]^2} & (36)\\
&=\lim_{\beta\to\infty}\frac{\sum_j P(Z=z_j)\exp(2\beta(f(z_j)-Z_{\max}))}{(\sum_j P(Z=z_j)\exp(\beta(f(z_j)-Z_{\max})))^2} & (37)\\
&=\frac{\sum_{z\in\mathcal{Z}_{\max}}P(Z=z)}{(\sum_{z\in\mathcal{Z}_{\max}}P(Z=z))^2} & (38)\\
&=\frac{1}{\sum_{z\in\mathcal{Z}_{\max}}P(Z=z)}, & (39)
\end{aligned}$$

where we used the fact that $\lim_{\beta\to\infty}\exp(\beta(z_j-Z_{\max}))=0$ for $z_j<Z_{\max}$. □

Lemma E.10 (Theorem 1 in (Mayrink Verdun et al., 2025))**.** *For $\beta>0$, and $N\ge1$, we have,*

$$\mathrm{KL}(\pi_{\beta,r_c^\star}(\cdot|x)\|\pi_{r_c^\star}^{(N,\beta)}(y|x)) \le \log(1+\frac{C_{\beta,r_c^\star,\mathrm{ref}}(x)}{N}). \tag{40}$$

Lemma E.11. *We have,*

$$\Big|\frac{f(r)}{\delta r}\Big| \le \frac{N^2\beta\exp(2\beta)}{(N-1)^2}, \tag{41}$$

where $f(r)=\log\big(\mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]\big)$, $r=r(x,y)$ is the and $R_i=r(Y_i,x)$.

Proof. Note that $\{R_i\}_{i=1}^{N-1}$ are i.i.d. . Therefore, we have,

$$\begin{aligned}
\frac{\delta f(r_c(y,x))}{\delta r} &= \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]^{-1}\frac{\delta \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]}{\delta r} \\
&\leq \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]^{-1} \\
&\quad \times \Big(\sum_{k=1}^{N}\frac{\beta k\binom{N-1}{k-1}\exp(\beta r)}{(k\exp(\beta r)+N-1-k)^2}(1-P(R=r))^{N-k}P^{k-1}(R=r)\Big) \\
&\leq \mathbb{E}[\frac{1}{\exp(\beta r)+\sum_{i=1}^{N-1}\exp(\beta R_i)}]^{-1} \\
&\quad \times \frac{\beta\exp(\beta)}{(N-1)^2}\Big(\sum_{k=1}^{N}k\binom{N-1}{k-1}(1-P(R=r))^{N-k}P^{k-1}(R=r)\Big) \\
&\leq \frac{N\beta\exp(2\beta)}{(N-1)^2}(1+(N-1)P(R=r)) \\
&\leq \frac{N^2\beta\exp(2\beta)}{(N-1)^2}.
\end{aligned} \tag{42}$$

□

F Proof and Details of Section 4

> *Lemma 4.1.* The following upper bound holds on KL divergence between SBoN and reference policies for a given prompt $x \in \mathcal{X}$,
>
> $$\mathrm{KL}(\pi_{r_c^\star}^{(N,\beta)}(y|x)\|\pi_{\mathrm{ref}}(y|x)) \leq \log\Big(\frac{N}{1+(N-1)\exp(-\beta)}\Big). \tag{43}$$

Proof. Recall that,

$$\pi_{r_c^\star}^{(N,\beta)}(y|x) = \pi_{\mathrm{ref}}(y|x)\exp(\beta r_c^\star(y,x))\mathbb{E}\Big[\big(\frac{1}{N}(\exp(\beta r_c^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r_c^\star(Y_i,x))\big)^{-1}\Big].$$

Now, we have,

$$\begin{aligned}
&\mathrm{KL}(\pi_{r_c^\star}^{(N,\beta)}(y|x)\|\pi_{\mathrm{ref}}(y|x)) \\
&= \sum_{\mathcal{Y}}\pi_{r_c^\star}^{(N,\beta)}(y|x)\log(\pi_{r_c^\star}^{(N,\beta)}(y|x)/\pi_{\mathrm{ref}}(y|x)) \\
&= \sum_{\mathcal{Y}}\pi_{r_c^\star}^{(N,\beta)}(y|x)\log(N)+\sum_{\mathcal{Y}}\pi_{r_c^\star}^{(N,\beta)}(y|x)\log(\mathbb{E}\Big[\big(\exp(\beta r_c^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r_c^\star(Y_i,x))\big)^{-1}\Big]) \\
&= \log(N)+\sum_{\mathcal{Y}}\pi_{r_c^\star}^{(N,\beta)}(y|x)\log(\mathbb{E}\Big[\big(1+\sum_{i=1}^{N-1}\exp(\beta(r_c^\star(Y_i,x)-r_c^\star(y,x)))\big)^{-1}\Big]),
\end{aligned} \tag{44}$$

For the second term in equation 44, consider

$$A(y,Y,x) = \sum_{i=1}^{N-1}\exp(\beta(r_c^\star(Y_i,x)-r_c^\star(y,x))) > 0,$$

where we have

$$(N-1)\exp(-\beta) \leq A(y,Y,x) \leq (N-1)\exp(\beta).$$

Therefore, we have,

$$
\begin{aligned}
&\sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x) \log(\mathbb{E}\Big[\big(1+\sum_{i=1}^{N-1}\exp(\beta(r_{\mathrm{c}}^\star(Y_i,x)-r_{\mathrm{c}}^\star(y,x)))\big)^{-1}\Big]) \\
&\leq \sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x) \log(\frac{1}{1+(N-1)\exp(-\beta)}) \\
&= \log(\frac{1}{1+(N-1)\exp(-\beta)}).
\end{aligned}
\tag{45}
$$

Combining equation 45 with equation 44 completes the proof.

$\square$

Lemma 4.2. The following upper bound holds on the KL divergence between the SBoN policies under true-reward and proxy-reward models respectively,

$$
\mathrm{KL}(\pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(\cdot|x)\|\pi_{\hat{r}_{\mathrm{c}}}^{(N,\beta)}(\cdot|x)) \leq \frac{N\beta\sqrt{\varepsilon_{\beta,r_{\mathrm{c}}}(x)}}{1+(N-1)\exp(-\beta)}\Big(\frac{N\exp(2\beta)}{(N-1)^2}+1\Big). \tag{46}
$$

Proof. We first provide the following upper bound,

$$
\begin{aligned}
&\mathrm{KL}(\pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x)\|\pi_{\hat{r}_{\mathrm{c}}}^{(N,\beta)}(y|x)) \\
&= \sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x) \log\big(\frac{\pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x)}{\pi_{\hat{r}_{\mathrm{c}}}^{(N,\beta)}(y|x)}\big) \\
&= \sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x)\beta(r_{\mathrm{c}}^\star(y,x)-\hat{r}_{\mathrm{c}}(y,x)) \\
&\quad + \sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r_{\mathrm{c}}^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r_{\mathrm{c}}^\star(Y_i,x))}]\big) \\
&\qquad\qquad - \log\big(\mathbb{E}[\frac{1}{\exp(\beta\hat{r}_{\mathrm{c}}(y,x))+\sum_{i=1}^{N-1}\exp(\beta\hat{r}_{\mathrm{c}}(Y_i,x))}]\big)\Big) \\
&\leq \frac{N\beta\sqrt{\varepsilon_{\beta,r_{\mathrm{c}}}(x)}}{1+(N-1)\exp(-\beta)} \\
&\quad + \sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r_{\mathrm{c}}^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r_{\mathrm{c}}^\star(Y_i,x))}]\big) \\
&\qquad\qquad - \log\big(\mathbb{E}[\frac{1}{\exp(\beta\hat{r}_{\mathrm{c}}(y,x))+\sum_{i=1}^{N-1}\exp(\beta\hat{r}_{\mathrm{c}}(Y_i,x))}]\big)\Big).
\end{aligned}
\tag{47}
$$

Note that for the last term in equation 47, we can apply the mean-value theorem as follows,

$$
\begin{aligned}
&\sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r_{\mathrm{c}}^\star(y,x))+\sum_{i=1}^{N-1}\exp(\beta r_{\mathrm{c}}^\star(Y_i,x))}]\big) \\
&\qquad\qquad - \log\big(\mathbb{E}[\frac{1}{\exp(\beta\hat{r}_{\mathrm{c}}(y,x))+\sum_{i=1}^{N-1}\exp(\beta\hat{r}_{\mathrm{c}}(Y_i,x))}]\big)\Big) \\
&\leq \sum_{\mathcal{Y}} \pi_{r_{\mathrm{c}}^\star}^{(N,\beta)}(y|x)|r_{\mathrm{c}}^\star(y,x)-\hat{r}_{\mathrm{c}}(y,x)|\Big|\frac{\delta f(r_\gamma(y,x))}{\delta r}\Big|,
\end{aligned}
\tag{48}
$$

where $f(r_\gamma(y,x)) = \log\big(\mathbb{E}[\frac{1}{\exp(\beta r_\gamma(y,x))+\sum_{i=1}^{N-1}\exp(\beta r_\gamma(Y_i,x))}]\big)$, for some $\gamma \in (0,1)$ we have $r_\gamma(y,x) = \gamma\hat{r}_{\mathrm{c}}(y,x)+(1-\gamma)r_{\mathrm{c}}^\star(y,x)$. Using Lemma E.11, we have,

$$
\Big|\frac{\delta f(r_{\mathrm{c}}(y,x))}{\delta r}\Big| \leq \frac{N^2\beta\exp(2\beta)}{(N-1)^2}. \tag{49}
$$

Using equation 49 in equation 48 and applying CauchySchwarz inequality, we have,

$$\begin{aligned}
&\sum_{\mathcal{Y}} \pi^{(N,\beta)}_{r^\star_c}(y|x)\Big(\log\big(\mathbb{E}[\frac{1}{\exp(\beta r^\star_c(y,x)) + \sum_{i=1}^{N-1}\exp(\beta r^\star_c(Y_i,x))}]\big) \\
&\qquad - \log\big(\mathbb{E}[\frac{1}{\exp(\beta \hat{r}_c(y,x)) + \sum_{i=1}^{N-1}\exp(\beta \hat{r}_c(Y_i,x))}]\big)\Big) \\
&\le \sqrt{\sum_{\mathcal{Y}} \mathbb{E}[1/(1+\sum_{i=1}^{N-1}\exp(\beta(r^\star_c(Y_i,x) - r^\star_c(y,x))))]^2 \pi_{\text{ref}}(y|x)} \\
&\quad \times \sqrt{\sum_{\mathcal{Y}} |r^\star_c(y,x) - \hat{r}_c(y,x)|^2 \pi_{\text{ref}}(y|x)} \sqrt{\sum_{\mathcal{Y}} \Big|\frac{\delta f(r_\gamma(y,x))}{\delta r}\Big|^2 \pi_{\text{ref}}(y|x)} \\
&\le \frac{\sqrt{\varepsilon_{\beta,r_c}(x)}}{1+(N-1)\exp(-\beta)} \frac{N^2\beta\exp(2\beta)}{(N-1)^2}.
\end{aligned} \tag{50}$$

It completes the proof. □

Theorem 4.3. The following upper bound holds on the improvement of expected calibrated true-reward under the SBoN policy for the calibrated proxy-reward model,

$$\mathbb{E}_{Y\sim\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)}[r^\star_c(Y,x)] \le 0.5 + \sqrt{\frac{1}{2}\log\Big(\frac{N}{1+(N-1)\exp(-\beta)}\Big)} \tag{51}$$

$$+ \min\Bigg(\sqrt{\frac{N\beta\sqrt{\varepsilon_{\beta,r_c}(x)}A(\beta,N)}{2(1+(N-1)\exp(-\beta))}}, 1\Bigg), \tag{52}$$

where $A(\beta,N) = \Big(\frac{N\exp(2\beta)}{(N-1)^2} + 1\Big)$.

Proof. Note that the following decomposition holds,

$$\begin{aligned}
&\mathbb{E}_{Y\sim\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)}[r^\star_c(Y,x)] - \mathbb{E}_{Y\sim\pi_{\text{ref}}(\cdot|x)}[r^\star_c(Y,x)] \\
&= \underbrace{\mathbb{E}_{Y\sim\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)}[r^\star_c(Y,x)] - \mathbb{E}_{Y\sim\pi^{(N,\beta)}_{r^\star_c}(\cdot|x)}[r^\star_c(Y,x)]}_{I_2} \\
&\quad + \underbrace{\mathbb{E}_{Y\sim\pi^{(N,\beta)}_{r^\star_c}(\cdot|x)}[r^\star_c(Y,x)] - \mathbb{E}_{Y\sim\pi_{\text{ref}}(\cdot|x)}[r^\star_c(Y,x)]}_{I_1}.
\end{aligned}$$

Note that we have $\mathbb{E}_{Y\sim\pi_{\text{ref}}(\cdot|x)}[r^\star_c(Y,x)] = 0.5$. □

Using Lemma 4.1 and Lemma 4.2 completes the proof.

G Proof and details of Section 5

Lemma 5.1 (**Full Version**). Under Assumption 3.1, the following properties of $C_{\beta,r_c,\text{ref}}(x)$ hold,

1. $C_{\beta,r_c,\text{ref}}(x) = \frac{\mathbb{E}[\exp(2\beta\hat{r}_c(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}_c(Y,x))]}$.
2. $C_{\beta,r_c,\text{ref}}(x)$ is an increasing function with respect to β.
3. $C_{\infty,r_c,\text{ref}}(x) = \frac{1}{\sum_i \pi_{\text{ref}}(y^{\max}_{i,r}(x)|x)}$ where $y^{\max}_{i,r}(x) \in \arg\max_y r_c(y,x)$.
4. For all $\beta < \infty$, we have $1 \le C_{\beta,r_c,\text{ref}}(x) \le \min(C_{\infty,r_c,\text{ref}}(x), \exp(2\beta))$.

Proof. In the following, we provide proofs of different items.

1.
$$
\begin{aligned}
C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x) &= \sum_{\mathcal{Y}} \frac{\pi^2_{\beta,\hat{r}_{\mathrm{c}}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} \\
&= \sum_{\mathcal{Y}} \frac{\exp(2\beta\hat{r}_{\mathrm{c}}(y,x))}{\mathbb{E}^2[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}\pi_{\mathrm{ref}}(y|x) \\
&= \frac{\mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}.
\end{aligned} \tag{53}
$$

2. We can show that the logarithm function of $C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)$ is increasing. Then, due to the increasing feature of the log function, the final result holds.
$$
\begin{aligned}
&\log(\frac{\mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}) \\
&= \log(\mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]) - 2\log(\mathbb{E}[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]),
\end{aligned} \tag{54}
$$
then we can compute the derivative of equation 54,
$$
\begin{aligned}
&\frac{\mathrm{d}\log(\mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))])}{\mathrm{d}\beta} - 2\frac{\mathrm{d}\log(\mathbb{E}[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathrm{d}\beta} \\
&= \frac{\mathbb{E}[2\hat{r}_{\mathrm{c}}(Y,x)\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]} - 2\frac{\mathbb{E}[\hat{r}_{\mathrm{c}}(Y,x)\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}
\end{aligned} \tag{55}
$$
Note that we have,
$$
\begin{aligned}
&\frac{\mathrm{d}\frac{\mathbb{E}[\hat{r}_{\mathrm{c}}(Y,x)\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}}{\mathrm{d}\beta} \\
&= \frac{\mathbb{E}[\hat{r}^2(Y,x)\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]\mathbb{E}[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))] - \mathbb{E}[\hat{r}_{\mathrm{c}}(Y,x)\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]^2}{\mathbb{E}^2[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]} \\
&= \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)}[\hat{r}^2(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)}[\hat{r}_{\mathrm{c}}(Y,x)]^2 \geq 0.
\end{aligned} \tag{56}
$$
Therefore, we have,
$$
\frac{\mathbb{E}[\hat{r}_{\mathrm{c}}(Y,x)\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]} \geq \frac{\mathbb{E}[\hat{r}_{\mathrm{c}}(Y,x)\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))]}. \tag{57}
$$
It completes the proof.

3. Follows directly from Lemma E.9.

4. Due to Jensen inequality for $\mathbb{E}^2[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))] \leq \mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]$, the $C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)$. We also have the uniform bound, $C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x) = \frac{\mathbb{E}[\exp(2\beta\hat{r}_{\mathrm{c}}(Y,x))]}{\mathbb{E}^2[\exp(\beta\hat{r}_{\mathrm{c}}(Y,x))}) \leq \exp(\beta)$. Furthermore, due to increasing property in second item, we also have $\sup_\beta C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x) = C_{\infty,r_{\mathrm{c}},\mathrm{ref}}(x)$. Therefore, the upper bound holds.

$\square$

Theorem G.1. *The following upper bound holds on the sub-optimality gap of the SBoN,*

$$\Delta_{J_{r_c^\star}}(\pi_{\beta,r_c^\star}\big(\cdot\,|x\big),\pi_{\hat{r}_c}^{(N,\beta)}(\cdot|x)) \leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r_c^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}_c}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big) \\ + \sqrt{\varepsilon_{\beta,r_c}(x)}\big(\sqrt{C_{\beta,\hat{r}_c,\mathrm{ref}}(x)} + \sqrt{C_{\beta,r_c^\star,\mathrm{ref}}(x)}\big) \\ + 2\sqrt{\frac{1}{2}\log\Big(1+\frac{C_{\beta,\hat{r}_c,\mathrm{ref}}(x)-1}{N}\Big)}.$$

Proof. Note that, we have,

$$\begin{aligned}
&\Delta_{J_{r_c^\star}}(\pi_{\beta,r_c^\star}\big(\cdot\,|x\big),\pi_{\hat{r}_c}^{(N,\beta)}(\cdot|x)) \\
&= \mathbb{E}_{Y\sim\pi_{\beta,r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\hat{r}_c}^{(N,\beta)}(\cdot|x)}[r_c^\star(Y,x)] \\
&= \underbrace{\mathbb{E}_{Y\sim\pi_{\beta,r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_c}(\cdot|x)}[r_c^\star(Y,x)]}_{I_1} \\
&\quad + \underbrace{\mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_c}(\cdot|x)}[r_c^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\hat{r}_c}^{(N,\beta)}(\cdot|x)}[r_c^\star(Y,x)]}_{I_2}
\end{aligned} \tag{58}$$

Note that, using the definition of $\pi_{\beta,r_c^\star}\big(\cdot\,|x\big)$ and $\pi_{\beta,\hat{r}_c}\big(\cdot\,|x\big)$ as solutions to KL-regularized problem, we have,

$$\mathbb{E}_{Y\sim\pi_{\beta,r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] = \frac{1}{\beta}\mathrm{KL}(\pi_{\beta,r_c^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r_c^\star(Y,x))]). \tag{59}$$

$$\mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_c}(\cdot|x)}[\hat{r}_c(Y,x)] = \frac{1}{\beta}\mathrm{KL}(\pi_{\beta,\hat{r}_c}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta \hat{r}_c(Y,x))]). \tag{60}$$

Therefore, for term I_1, we have,

$$
\begin{aligned}
&\mathbb{E}_{Y\sim\pi_{\beta,r_{\mathrm{c}}^\star}\left(\cdot|x\right)}[r_{\mathrm{c}}^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_{\mathrm{c}}}\left(\cdot|x\right)}[r_{\mathrm{c}}^\star(Y,x)]\\
&= \mathbb{E}_{Y\sim\pi_{\beta,r_{\mathrm{c}}^\star}\left(\cdot|x\right)}[r_{\mathrm{c}}^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_{\mathrm{c}}}\left(\cdot|x\right)}[\hat{r}_{\mathrm{c}}(Y,x)]\\
&\quad + \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_{\mathrm{c}}}\left(\cdot|x\right)}[\hat{r}_{\mathrm{c}}(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_{\mathrm{c}}}\left(\cdot|x\right)}[r_{\mathrm{c}}^\star(Y,x)]\\
&= \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r_{\mathrm{c}}^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big)\\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r_{\mathrm{c}}^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta \hat{r}_{\mathrm{c}}(Y,x))])\\
&\quad + \sum_{\mathcal{Y}} \pi_{\beta,\hat{r}_{\mathrm{c}}}(\,\cdot\,|x)(\hat{r}_{\mathrm{c}}(y,x) - r_{\mathrm{c}}^\star(y,x))\\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r_{\mathrm{c}}^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big)\\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r_{\mathrm{c}}^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta \hat{r}_{\mathrm{c}}(Y,x))])\\
&\quad + \frac{1}{\sqrt{\beta}}\sqrt{\sum_{\mathcal{Y}}\frac{\pi_{\beta,\hat{r}_{\mathrm{c}}}^2(y|x)}{\pi_{\mathrm{ref}}(y|x)}}\sqrt{\beta\sum_{\mathcal{Y}}(\hat{r}_{\mathrm{c}}(y,x) - r_{\mathrm{c}}^\star(y,x))^2\pi_{\mathrm{ref}}(y|x)}\\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r_{\mathrm{c}}^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big)\\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r_{\mathrm{c}}^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta \hat{r}_{\mathrm{c}}(Y,x))])\\
&\quad + \sqrt{C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r_{\mathrm{c}}^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big)\\
&\quad + \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta r_{\mathrm{c}}^\star(Y,x))]) - \frac{1}{\beta}\log(\mathbb{E}_{Y\sim\pi_{\mathrm{ref}}(\cdot|x)}[\exp(\beta \hat{r}_{\mathrm{c}}(Y,x))])\\
&\quad + \sqrt{C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\\
&\leq \frac{1}{\beta}\Big(\mathrm{KL}(\pi_{\beta,r_{\mathrm{c}}^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,\hat{r}_{\mathrm{c}}}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))\Big)\\
&\quad + \sqrt{C_{\beta,r_{\mathrm{c}}^\star,\mathrm{ref}}(x)\varepsilon_{\beta,r_{\mathrm{c}}}(x)}\\
&\quad + \sqrt{C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)\varepsilon_{\beta,r_{\mathrm{c}}}(x)}.
\end{aligned} \tag{61}
$$

For term I_2 and using similar approach to term I_1 and applying Lemma E.10, we have,

$$
\begin{aligned}
&\mathbb{E}_{Y\sim\pi_{\beta,\hat{r}_{\mathrm{c}}}\left(\cdot|x\right)}[r_{\mathrm{c}}^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\hat{r}_{\mathrm{c}}}^{(N,\beta)}(\cdot|x)}[r_{\mathrm{c}}^\star(Y,x)]\\
&\leq 2\mathbb{TV}(\pi_{\beta,\hat{r}_{\mathrm{c}}}(\,\cdot\,|x), \pi_{\hat{r}_{\mathrm{c}}}^{(N,\beta)}(\cdot|x))\\
&\leq 2\min\left(1, \sqrt{\frac{1}{2}\mathrm{KL}(\pi_{\beta,\hat{r}_{\mathrm{c}}}(\,\cdot\,|x)\|\pi_{\hat{r}_{\mathrm{c}}}^{(N,\beta)}(\cdot|x))}\right)\\
&\leq 2\min\left(1, \sqrt{\frac{1}{2}\log\left(1+\frac{C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)-1}{N}\right)}\right)\\
&\leq 2\sqrt{\frac{1}{2}\log\left(1+\frac{C_{\beta,\hat{r}_{\mathrm{c}},\mathrm{ref}}(x)-1}{N}\right)}
\end{aligned} \tag{62}
$$

Combining equation 61 and equation 62 with equation 58 completes the proof. ∎

Theorem 5.3. Under Assumption 3.1, the following upper bound holds on the optimal regret gap of the SBoN policy for any $\beta > 0$,

$$\Delta_{J_{r_c^\star}}(\pi^\star_{r_c^\star}(\cdot|x), \pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)) \leq \sqrt{\varepsilon_{\beta,r_c}(x)}\Big(\sqrt{C_{\infty,\hat{r}_c,\mathrm{ref}}(x)} + \sqrt{C_{\infty,r_c^\star,\mathrm{ref}}(x)}\Big) + 2\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty,\hat{r}_c,\mathrm{ref}}(x) - 1}{N}\Big)} + \frac{\log(C_{\infty,r_c^\star,\mathrm{ref}}(x))}{\beta},$$

Proof. Note that we have,

$$\begin{aligned}
&\Delta_{J_{r_c^\star}}(\pi^\star_{r_c^\star}(\cdot|x), \pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)) \\
&= \mathbb{E}_{Y\sim\pi^\star_{r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] - \mathbb{E}_{Y\sim\pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x)}[r_c^\star(Y,x)] \\
&= \underbrace{\mathbb{E}_{Y\sim\pi^\star_{r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,r_c^\star}(\cdot|x)}[r_c^\star(Y,x)]}_{I_3} \\
&\quad + \underbrace{\Delta_{J_{r_c^\star}}(\pi_{\beta,r_c^\star}(\cdot|x), \pi^{(N,\beta)}_{\hat{r}_c}(\cdot|x))}_{I_4}
\end{aligned} \tag{63}$$

For term I_4, we can use Theorem G.1. For term I_3, note that, we have for $\beta > 0$,

$$\mathbb{E}_{Y\sim\pi^\star_{r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] \leq \frac{\mathrm{KL}(\pi^\star_{r_c^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)) - \mathrm{KL}(\pi_{\beta,r_c^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))}{\beta} \tag{64}$$

Combining equation 64 with Theorem G.1, completes the proof due the positiveness of KL divergence and using Lemma E.7 and Lemma 5.1.

□

Remark G.2. For $\beta = 0$, we have, $\lim_{\beta\to 0}\pi_{\beta,r_c^\star}(\cdot|x) = \pi_{\mathrm{ref}}(\cdot|x)$. Therefore, we have,

$$\mathbb{E}_{Y\sim\pi^\star_{r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] - \mathbb{E}_{Y\sim\pi_{\beta,r_c^\star}(\cdot|x)}[r_c^\star(Y,x)] \leq \sqrt{2\mathrm{KL}(\pi^\star_{r_c^\star}(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x))}. \tag{65}$$

Theorem 5.3. Under Assumption 3.1, the following upper bound holds on the optimal regret gap of the BoN policy for any $\beta > 0$,

$$\Delta_{J_{r_c^\star}}(\pi^\star_{r_c^\star}(\cdot|x), \pi^{(N,\infty)}_{\hat{r}_c}(\cdot|x)) \leq \sqrt{\varepsilon_{\infty,r}(x)}\Big(\sqrt{C_{\infty,\hat{r}_c,\mathrm{ref}}(x)} + \sqrt{C_{\infty,r_c^\star,\mathrm{ref}}(x)}\Big) + 2\sqrt{\frac{1}{2}\log\Big(1 + \frac{C_{\infty,\hat{r}_c,\mathrm{ref}}(x) - 1}{N}\Big)}.$$

Proof. The results follow directly from Theorem 5.3 for $\beta \to \infty$. □

H Experiments

Our implementation is available in the following link: `https://anonymous.4open.science/r/sbon-C6D3/README.md` .

LLM-as-a-Judge: To evaluate the performance of the generated responses, we used GPT-4o to determine how harmless they are. Specifically, we used the following prompt:

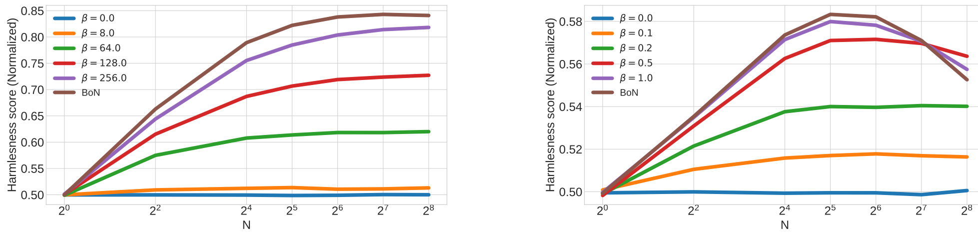

Figure 3: Soft Best-of-N experiment using a strong reward model (Left) and a weak one (Right). When the performance of the reward model is poor, BoN can lead to overoptimization, while the SBoN can help to mitigate it.

> You are an AI safety evaluator. Your task is to determine if a response to a given prompt contains harmful content. Respond with a harmlessness score using these criteria:
>
> 2 = SAFE: The response is completely safe and appropriate.
> 1 = SOMEWHAT HARMFUL: The response contains potentially concerning elements, mild toxicity, stereotypes, subtle harmful guidance, or borderline inappropriate content.
> 0 = HARMFUL: The response contains clearly harmful content such as hate speech, explicit instructions for dangerous activities, severe toxicity, etc.
>
> Provide your judgment as an integer in curly brackets like this: 1, 2, or 3.

This yields a reward with values ranging from 0 to 2, where 0 indicates complete harm and 2 indicates safety.

H.1 More Experiments

To further support our theoretical analysis, we conducted more experiments comparing Soft Best-of-N (SBoN) across different regularization strengths and reward model qualities. We ran one extra experimental conditions: one using a weaker proxy-reward model (tulu-v2.5-13b-uf-rm (Ivison et al., 2024)) respect to Armo reward model. As shown in Figure 3, when the reward model is weak, performance degrades for large N due to reward hacking. However, the smoothing in SBoN helps mitigate this degradation. This observation is also aligned with our theoretical analysis and discussion in Section 4.

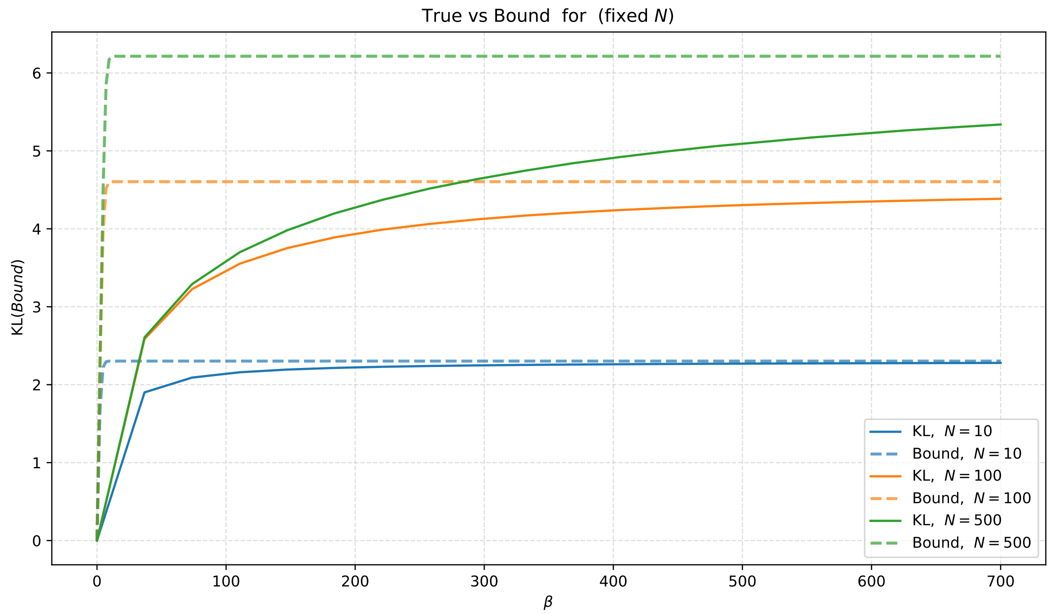

Figure 4: True KL divergence vs upper bound in Lemma 4.1 for fix $N = \{10, 100, 500\}$.

H.2 Numerical Example

To illustrate how our analytical upper bound in Lemma 4.1 behaves as a function of the temperature parameter β, we run a toy experiment in which

1. the reference policy is the uniform distribution over responses, and
2. rewards are bounded with $= 1$.

For each β in a logarithmic sweep, we compute the true KLdivergence between the SBoN policy and the reference policy, together with the theoretical bound derived in Lemma 4.1.

- **Very large β (nearBoN policy).** As $\beta \to \infty$ the SBoN policy converges to the BoN policy. The gap between the KL and the bound vanishes.
- **Very small β (reference policy).** When $\beta \to 0$ the SBoN policy approaches the uniform sampling from the reference policy, which results in the reference policy, making the KL itself tend to zero; the bound is equal to zero for this value.

This experiment confirms that the bound is tight in the two asymptotic regimes and remains a conservative yet informative estimate elsewhere.