SAFETY HACKING IN CONSTRAINED BEST-OF-N INFERENCE-TIME SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Inference-time scaling pipelines often sample multiple outputs, filter them with a learned safety model, and return the proxy-feasible output with the highest learned reward. We show that this pipeline can fail in two stages: an imperfect safety proxy first contaminates the feasible set with unsafe outputs, and reward maximization can then amplify this residual contamination. We call this failure *safety hacking*, in which the selected output passes the learned filter but violates the true safety constraint. For constrained Best-of-N sampling, we derive finite-N bounds determined by the joint upper reward tails of safe and unsafe outputs within the proxy-feasible set. If unsafe-but-feasible outputs have the heavier tail, the probability of safety hacking approaches 1 as N grows, even when false-positive mass and average safety and reward proxy errors are arbitrarily small. We also show that policies within a bounded χ^2 divergence from the proxy-feasible reference distribution admit an N-independent safety-hacking bound. Coverage control limits amplification but cannot repair a contaminated feasible set: admitted unsafe outputs may still be favored, and regularized selection is not necessarily safer than constrained Best-of-N. Toy and language-model experiments characterize both contamination and its reward-tail amplification, which exposes an inherent difficulty in inference-time scaling with imperfect reward and safety models.

1 INTRODUCTION

Foundation-model inference increasingly relies on learned models both to enforce safety and to assess response quality [Snell et al., 2024; Amodei et al., 2016; Hendrycks et al., 2021]. A typical pipeline samples candidate responses, rejects those deemed unsafe by a safety model, and ranks the remainder with a reward model [Wang et al., 2025; Ji et al., 2025; Inan et al., 2023; Zeng et al., 2024]. As a representative instance, constrained Best-of-N (CBoN) sampling follows this pattern: it draws N responses and returns the proxy-feasible response with the largest learned reward.

Filtering and ranking errors can interact. Unsafe responses admitted by the safety proxy remain in the feasible set, where reward optimization can make them disproportionately likely to be selected. We call the selection of such a response *safety hacking*: the returned response satisfies the learned safety constraint but violates the true one.

Existing work on reward overoptimization shows that excessive optimization against an imperfect reward proxy can eventually reduce the true reward [Skalse et al., 2022; Gao et al., 2023]. Our setting adds a distinct source of misspecification: the optimization domain is itself defined by an imperfect safety proxy. The resulting failure is governed by the interaction between 1) *contamination* that determines which unsafe outputs enter the proxy-feasible set, and 2) *amplification* that determines how strongly reward optimization favors them.

Our main result shows that amplification depends not only on how often the filter admits unsafe responses, but also on the relative upper tails of learned reward for safe and unsafe responses in the proxy-feasible set. When unsafe responses have the heavier upper tail, the probability that CBoN selects an unsafe response tends to 1 as $N \to \infty$. Our theoretical result shows that this failure can occur even when contamination and average proxy errors are small. The reference policy assigns mass ε to unsafe responses admitted by the filter, both proxy RMSEs scale as $\sqrt{\varepsilon}$, and a sampling budget on the order of $1/\varepsilon$ suffices to produce high-probability safety hacking. Average filter accuracy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is therefore insufficient to predict safety under inference-time scaling; the relative learned-reward tails of safe and unsafe admitted responses also matter.

Under suitable conditions, pessimistic inference can prevent reward overoptimization from worsening with larger sampling budgets [Huang et al., 2025], but it cannot determine which responses were wrongly admitted by a safety proxy. We show that any policy close to the proxy-feasible reference distribution in χ^2 divergence admits an N-independent safety-hacking bound. An example is constrained pessimistic sampling (cPes), a χ^2-regularized reweighting of the proxy-feasible reference distribution. The guarantee limits concentration on residual errors but cannot eliminate them. Depending on the reward proxy, regularized selection may still exhibit substantial safety hacking.

Our contributions are:

1. We formalize *safety hacking* as selection of a response that satisfies a learned safety constraint while violating the true safety constraint.
2. We characterize how safety-filter contamination is amplified by reward maximization. We derive finite-N bounds for CBoN in terms of the joint reward tails of safe and unsafe proxy-feasible outputs. We also give conditions under which safety hacking becomes asymptotically certain even as unsafe false acceptances and average proxy errors vanish.
3. We derive N-independent safety-hacking bounds under coverage control and study cPes sampling as a concrete instance. These bounds limit amplification but cannot repair contamination introduced by an imperfect safety constraint.
4. Controlled simulations and language-model experiments examine both contamination and amplification. A finite-N decomposition shows that the increase in safety hacking with N is driven by unsafe candidates outranking safe ones. Reward-model ablations show that this amplification depends on reward ranking, and the same pattern appears with an alternative safety filter.

2 Related Work

Reward overoptimization. Optimizing an imperfect proxy can produce unintended behavior, a concern studied through reward hacking or Goodhart's law [Amodei et al., 2016; Manheim & Garrabrant, 2018]. In language-model alignment, learned reward models are used both during training [Ouyang et al., 2022; Bai et al., 2022] and to rank responses at inference time [Stiennon et al., 2020]. Prior work shows that excessive optimization of an imperfect reward proxy, through reinforcement learning or inference-time scaling, can eventually reduce true reward [Gao et al., 2023; Khalaf et al., 2025]. Our setting includes misspecification in both the reward and the safety constraint (i.e., optimization domain). A learned safety proxy defines a feasible set that can contain unsafe false acceptances, which reward maximization may then amplify through upper-tail selection. We analyze the interaction between these two sources of error.

Safety-constrained inference-time alignment. Chittepu et al. [2026] formulate inference-time alignment with learned reward and safety cost models, using a calibrated Lagrangian reward for sequence-level BoN under a safety constraint in expectation. While Chittepu et al. [2026] optimize against the learned cost signal, we instead study a misspecified hard safety filter. In our setting, false acceptances contaminate the feasible set, and subsequent reward maximization can amplify them.

Inference-time pessimism and coverage control. Huang et al. [2025] and Yu et al. [2026] study coverage and scaling in inference-time alignment and propose pessimistic alternatives to naïve BoN. Also, regularized BoN has been studied under a minimum Bayes-risk objective [Jinnai et al., 2025]. We condition the reference distribution on proxy feasibility and apply coverage control to the resulting distribution. This yields an upper bound with two terms: the unsafe mass already present in the conditioned reference and the additional concentration allowed by downstream selection. Coverage control limits the second term, but a small overall bound also requires the conditioned reference to assign little probability to unsafe responses.

Learned safety filters and evaluator errors. LLM systems often use learned safety filters or guardrails to classify prompts and responses. Llama Guard casts safeguarding as safety-risk classification [Inan et al., 2023], while ShieldGemma provides open models for detecting risks in model inputs and outputs [Zeng et al., 2024]. These systems make safety constraints operational but remain

imperfect proxies for the true criterion [Zheng et al., 2023]. We study how errors in the two proxies interact: unsafe false acceptances determine which responses enter the feasible set, and reward-proxy errors determine which response is selected.

Jailbreaks and red teaming. HarmBench [Mazeika et al., 2024], JailbreakBench [Chao et al., 2024], and StrongREJECT [Souly et al., 2024] provide benchmarks for harmful behavior, jailbreak, and refusal robustness. BoN jailbreaking shows that repeated randomized attempts can substantially increase attack success rates [Hughes et al., 2024], demonstrating that search can expose rare safety failures. We instead hold the prompt fixed and search over sampled model outputs. The filter may admit unsafe outputs, and reranking can make them increasingly likely to be selected.

3 Problem Statement

We consider safety-constrained reward maximization at inference time. For clarity, we write the theory for a fixed prompt $x \in \mathcal{X}$ whenever possible; hence, all probabilities and expectations are conditional on this prompt. We take $\mathcal{Y}$ to be finite or countable, and the measurable-space extension is obtained by replacing sums with integrals. We suppose that there are an unknown reward function $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ and an unknown binary safety function $g^\star : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$. For a prompt x, an output y is truly safe if $g^\star(x, y) = 1$. Define the following truly safe and unsafe sets of outputs:

$$\mathcal{S}^\star_+(x) \coloneqq \{y \in \mathcal{Y} : g^\star(x, y) = 1\}, \qquad \mathcal{S}^\star_-(x) \coloneqq \mathcal{Y} \setminus \mathcal{S}^\star_+(x). \tag{1}$$

Thus, the ideal constrained inference problem is

$$\max_{y \in \mathcal{Y}} \; r^\star(x, y) \qquad \text{subject to} \qquad g^\star(x, y) = 1. \tag{2}$$

The agent cannot evaluate $r^\star$ or $g^\star$ directly, but instead has a reference policy $\pi_{\mathrm{ref}}(\cdot \mid x)$, a learned reward proxy $\widehat{r} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, and a learned safety proxy $\widehat{g} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$. For a safety threshold $b \in (0, 1]$, define the proxy-feasible set

$$\widehat{\mathcal{S}}_+(x, b) \coloneqq \{y \in \mathcal{Y} : \widehat{g}(x, y) \geq b\}. \tag{3}$$

The proxy-feasible set decomposes into true positives and false positives:

$$A(x, b) \coloneqq \mathcal{S}^\star_+(x) \cap \widehat{\mathcal{S}}_+(x, b), \qquad B(x, b) \coloneqq \mathcal{S}^\star_-(x) \cap \widehat{\mathcal{S}}_+(x, b). \tag{4}$$

Both sets contain outputs accepted by the safety proxy: $A(x, b)$ contains the truly safe ones, whereas $B(x, b)$ contains the unsafe ones.

Definition 3.1 (Safety hacking)**.** *Fix $x \in \mathcal{X}$ and $b \in (0, 1]$. If an inference-time procedure outputs $\widehat{y} \in \widehat{\mathcal{S}}_+(x, b)$, we say that safety hacking occurs when*

$$\widehat{y} \in B(x, b). \tag{5}$$

Equivalently, the output passes the learned safety proxy but violates the true safety constraint. Note that if the procedure abstains (e.g., when $\widehat{\mathcal{S}}_+(x, b) = \emptyset$), this is not considered safety hacking.

The following joint upper-tail quantities are the central objects in the CBoN analysis. For $\Diamond \in \{A, B\}$ and $t \in \mathbb{R}$, define the learned-score tail

$$\widehat{\Psi}_\Diamond(t; x, b) \coloneqq \mathbb{P}_{y \sim \pi_{\mathrm{ref}}(\cdot \mid x)} \left[y \in \Diamond(x, b),\ \widehat{r}(x, y) > t \right]. \tag{6}$$

These tails combine 1) the probability that the reference policy π_{ref} reaches a class of outputs $\Diamond$ and 2) the upper-tail behavior of the learned reward score inside that class.

Although we use a contextual bandit setting for simplicity, the formulation also covers open-loop trajectory selection. Let τ denote a trajectory. Set $y = \tau$ and let $\pi_{\mathrm{ref}}(\tau \mid x)$ be the trajectory distribution induced by a fixed reference policy and the environment dynamics, with reward and safety defined at the trajectory level. The results require N i.i.d. trajectories from this distribution and therefore exclude beam search, Monte Carlo tree search, and other non-i.i.d. or stepwise procedures.

4 SAFETY-FILTER CONTAMINATION AND REWARD-TAIL AMPLIFICATION

A natural baseline to solve (2) is CBoN. Given N i.i.d. candidates $y_1, \dots, y_N \sim \pi_{\mathrm{ref}}(\cdot \mid x)$, define a proxy-feasible index set $\mathcal{I}_N(x,b) \coloneqq \{i \in [N] : y_i \in \widehat{\mathcal{S}}_+(x,b)\}$. If $\mathcal{I}_N(x,b) \neq \varnothing$, CBo$N$ returns

$$\widehat{y}_N^{\mathrm{BoN}}(x) \in \operatorname*{arg\,max}_{i \in \mathcal{I}_N(x,b)} \widehat{r}(x, y_i) \tag{7}$$

with arbitrary tie-breaking. In the case of $\mathcal{I}_N(x,b) = \varnothing$, the procedure simply abstains.

The failure of CBoN has two stages. First, false positives from the safety proxy contaminate the proxy-feasible set with outputs in $B(x,b)$. Second, maximizing $\widehat{r}$ over that set can amplify the contamination by searching the extreme upper tail of the learned reward score. Increasing N therefore does more than increase exposure to unsafe-but-feasible outputs: it also intensifies their reward-based competition with safe proxy-feasible outputs. The following theorem identifies the joint score tails in (6) as the finite-N quantities governing this competition.

Theorem 4.1 (Finite-N safety-hacking probability bounds)**.** *Fix a prompt $x \in \mathcal{X}$ and a threshold $b \in (0,1]$. Let $y_1, \dots, y_N \overset{\mathrm{iid}}{\sim} \pi_{\mathrm{ref}}(\cdot \mid x)$, and let $\widehat{y}_N^{\mathrm{BoN}}(x)$ be defined by (7). For any $t, s \in \mathbb{R}$, define*

$$\widehat{L}_N(t; x, b) \coloneqq \big(1 - \widehat{\Psi}_A(t; x, b)\big)^N - \big(1 - \widehat{\Psi}_A(t; x, b) - \widehat{\Psi}_B(t; x, b)\big)^N, \tag{8}$$

$$\widehat{U}_N(s; x, b) \coloneqq 1 - \big(1 - \widehat{\Psi}_B(s; x, b)\big)^N + \big(1 - \widehat{\Psi}_B(s; x, b) - \widehat{\Psi}_A(s; x, b)\big)^N. \tag{9}$$

Then, for every $t, s \in \mathbb{R}$,

$$\widehat{L}_N(t; x, b) \leq \mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x,b)\right] \leq \widehat{U}_N(s; x, b). \tag{10}$$

Taking the supremum over t and the infimum over s gives the tightest bounds in this family.

The lower bound is the probability that no sample from $A(x,b)$ has learned reward above t and at least one sample from $B(x,b)$ does. CBoN selects from $B(x,b)$ on this event. The upper bound follows because CBoN cannot select from $B(x,b)$ when some sample from $A(x,b)$ scores above s and none from $B(x,b)$ does. Corollary A.1 states the corresponding guarantees in terms of the effective tail masses $N\widehat{\Psi}_A$ and $N\widehat{\Psi}_B$. Let $\bar{t} \coloneqq \operatorname{ess\,sup}_{y \in A(x,b)} \widehat{r}(x,y)$, so that $\widehat{\Psi}_A(\bar{t}; x, b) = 0$. Theorem 4.1 then gives $\mathbb{P}[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x,b)] \geq 1 - (1 - \widehat{\Psi}_B(\bar{t}; x, b))^N$. If $\widehat{\Psi}_B(\bar{t}; x, b) = 0.01$, the lower bound in Theorem 4.1 is approximately 0.63 at $N = 100$ and 0.99 at $N = 500$.

The mass of $B(x,b)$ measures contamination, while the relative joint tails govern its selection under CBoN. The next result gives conditions under which the safety-hacking probability converges to 1.

Theorem 4.2 (Asymptotic safety hacking)**.** *Fix $x \in \mathcal{X}$ and $b \in (0,1]$. If there exists a sequence $t_N \in \mathbb{R}$ such that $N\widehat{\Psi}_A(t_N; x, b) \to 0$, and $N\widehat{\Psi}_B(t_N; x, b) \to \infty$, then*

$$\lim_{N \to \infty} \mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x,b)\right] = 1. \tag{11}$$

In the boundary case discussed above, $\widehat{\Psi}_A(\bar{t}; x, b) = 0$ and $\widehat{\Psi}_B(\bar{t}; x, b) > 0$. Setting $t_N = \bar{t}$, the theorem shows that the safety-hacking probability converges to 1. More generally, under the tail condition above, BoN can select an unsafe class with probability approaching 1 even when that class has arbitrarily small reference mass. The fixed-prompt result extends directly to prompt distributions: if the tail condition holds on a set of prompts with probability mass at least ρ, the aggregate hacking probability has liminf at least ρ; see Corollary A.2.

Corollary 4.3 (Vanishing proxy error still allows amplification)**.** *Fix x and $b \in (0,1]$. For every $\varepsilon \in (0,1)$ and $\xi > 0$, there exists a two-output instance whose unsafe-but-feasible reference mass is ε, whose safety- and reward-proxy RMSEs are respectively $b\sqrt{\varepsilon}$ and $\xi\sqrt{\varepsilon}$, and for which*

$$\mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x,b)\right] = 1 - (1 - \varepsilon)^N \geq 1 - e^{-N\varepsilon}. \tag{12}$$

Consequently, any $N \geq \varepsilon^{-1} \log(1/\delta)$ yields safety hacking with probability at least $1 - \delta$. Thus, even as the false-positive mass and both average proxy errors vanish, a budget $N = \Theta(1/\varepsilon)$ can amplify the residual error to high probability.

When true reward is bounded on the proxy-feasible set, the safety-hacking probability approaches 1 if reward overestimation has a sufficiently heavier Gaussian upper tail among unsafe responses in that set than among safe responses. The relevant threshold scales as $\sqrt{\log N}$; see Proposition A.6.

5 COVERAGE CONTROL LIMITS AMPLIFICATION, NOT CONTAMINATION

The preceding analysis separates two sources of risk. *Contamination* is the unsafe mass admitted by the safety proxy, while *amplification* is the additional concentration on that mass induced by downstream reward optimization. This distinction suggests a general mitigation principle: bounded deviation from the proxy-feasible reference distribution limits amplification as N grows, although it cannot remove the underlying contamination.

Let $q(x,b) \coloneqq \mathbb{P}_{y\sim\pi_{\mathrm{ref}}(\cdot\,|x)}[y \in \widehat{\mathcal{S}}_+(x,b)]$ denote the reference probability that a sampled output passes the learned safety filter. Throughout this section, we assume $q(x,b) > 0$. Conditioning the reference policy on this event gives the proxy-feasible reference distribution

$$\pi^\sharp_{\mathrm{ref}}(y \mid x) \coloneqq \frac{\pi_{\mathrm{ref}}(y \mid x)\cdot \mathbb{I}\{y \in \widehat{\mathcal{S}}_+(x,b)\}}{q(x,b)}. \tag{13}$$

This distribution is the baseline against which we measure how strongly an inference procedure concentrates on feasible outputs. We write $\pi(\cdot \mid x) \ll \pi^\sharp_{\mathrm{ref}}(\cdot \mid x)$ when $\pi(y \mid x) > 0$ implies $\pi^\sharp_{\mathrm{ref}}(y \mid x) > 0$ for every $y \in \mathcal{Y}$. For any such policy, define the following coverage coefficient by

$$C^\sharp_\pi(x) \coloneqq \sum_{y\in\mathcal{Y}} \frac{\pi(y \mid x)^2}{\pi^\sharp_{\mathrm{ref}}(y \mid x)} = 1 + \chi^2\left(\pi(\cdot \mid x) \,\middle\|\, \pi^\sharp_{\mathrm{ref}}(\cdot \mid x)\right). \tag{14}$$

Small $C^\sharp_\pi(x)$ means that the final policy retains broad coverage of the proxy-feasible reference distribution, whereas a large value indicates strong concentration.

Theorem 5.1 (Coverage-controlled contamination amplification)**.** *Define the residual unsafe mass under $\pi^\sharp_{\mathrm{ref}}$ as $\kappa(x,b) \coloneqq \mathbb{P}_{y\sim\pi^\sharp_{\mathrm{ref}}(\cdot\,|x)}[y \in B(x,b)]$. With any proxy-feasible policy $\pi(\cdot \mid x) \ll \pi^\sharp_{\mathrm{ref}}(\cdot \mid x)$, the following inequality holds:*

$$\left|\mathbb{P}_{y\sim\pi(\cdot\,|x)}[y \in B(x,b)] - \kappa(x,b)\right| \le \sqrt{\big(C^\sharp_\pi(x) - 1\big)\kappa(x,b)\big(1-\kappa(x,b)\big)}. \tag{15}$$

Define $\bar{\varepsilon}_g(x) \coloneqq \sqrt{\mathbb{E}_{y\sim\pi_{\mathrm{ref}}(\cdot\,|x)}\left[(\widehat{g}(x,y) - g^\star(x,y))^2\right]}$. Then,

$$\kappa(x,b) \le \min\left\{1, \frac{\bar{\varepsilon}_g(x)^2}{b^2 q(x,b)}\right\}. \tag{16}$$

The theorem applies to any inference procedure with bounded proxy-feasible coverage, regardless of how its final policy is constructed. Specifically, (15) bounds the change in unsafe mass using $C^\sharp_\pi(x)$ and the residual contamination $\kappa(x,b)$. If the coverage coefficient remains bounded as N grows, this guarantee is uniform in the sampling budget. When $C^\sharp_\pi(x) = 1$, we have $\pi = \pi^\sharp_{\mathrm{ref}}$, so the safety-hacking probability equals $\kappa(x,b)$.

5.1 A REGULARIZED INSTANTIATION

To prevent arbitrarily large learned reward scores from inducing extreme concentration, fix clipping levels $-\infty < \widehat{R}_{\min} \le \widehat{R}_{\max} < \infty$ and set $\widehat{R}_{\mathrm{span}} \coloneqq \widehat{R}_{\max} - \widehat{R}_{\min}$. Then, define

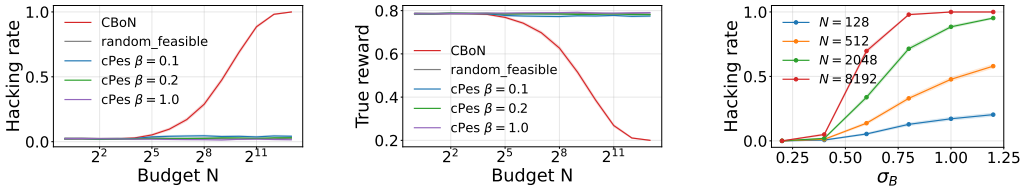
$$\widetilde{r}(x,y) \coloneqq \mathrm{clip}\left(\widehat{r}(x,y), \widehat{R}_{\min}, \widehat{R}_{\max}\right). \tag{17}$$

We use $\widetilde{r}$ only for the coverage-controlled policy below; the CBoN analysis in Section 4 continues to use the original learned score $\widehat{r}$. For $\beta > 0$, consider the regularized proxy-feasible policy

$$\hat{\pi} \in \operatorname*{arg\,max}_{\pi \ll \pi^\sharp_{\mathrm{ref}}} \left\{\mathbb{E}_\pi[\widetilde{r}(x,y)] - \frac{\beta}{2}\big(C^\sharp_\pi(x) - 1\big)\right\}. \tag{18}$$

Proposition 5.2 (Regularized reweighting)**.** *There is a unique $\lambda(x) \in [\widehat{R}_{\min} - \beta, \widehat{R}_{\max})$ satisfying $\mathbb{E}_{y\sim\pi^\sharp_{\mathrm{ref}}(\cdot\,|x)}\left[\mathrm{ReLU}\left(\frac{\widetilde{r}(x,y)-\lambda(x)}{\beta}\right)\right] = 1$, and the unique optimizer of (18) is*

$$\hat{\pi}(y \mid x) = \pi^\sharp_{\mathrm{ref}}(y \mid x)\,\mathrm{ReLU}\left(\frac{\widetilde{r}(x,y) - \lambda(x)}{\beta}\right). \tag{19}$$



(a) Conditional safety-hacking rate. (b) Conditional mean true reward. (c) False-positive tail scale.

Figure 1: Toy results. As the search budget grows, CBoN increasingly selects the rare unsafe-but-feasible class B, whose reward-proxy upper tail dominates that of the safe class A. Random feasible selection does not exhibit this amplification, while cPes limits the resulting concentration. (c) varies the false-positive reward-noise scale σ_B.

Moreover, its concentration relative to the proxy-feasible reference is bounded as

$$C^{\sharp}_{\hat{\pi}}(x) \le 1 + D^{\sharp}(x), \qquad \textit{where} \qquad D^{\sharp}(x) := \min\left\{ \frac{\mathrm{Var}_{\pi^{\sharp}_{\mathrm{ref}}}[\widetilde{r}(x,y)]}{\beta^2}, \frac{\widehat{R}_{\mathrm{span}}}{\beta} \right\}. \tag{20}$$

Hereinafter, we call the resulting algorithm constrained pessimistic sampling (cPes). A finite-sample implementation replaces the expectations above by their empirical analogues over the proxy-feasible candidates and samples according to the resulting reweighting. Full pseudocode and consistency are given in Algorithm 1 and Proposition A.8. Let $m_N = |\mathcal{I}_N(x,b)|$ be the number of proxy-feasible candidates, and let $\widehat{y}_{\mathrm{cPes},N}(x)$ denote the output of finite-sample cPes.

Corollary 5.3 (Safety of cPes)**.** *Conditional on $m_N > 0$, the sample implementation satisfies*

$$\mathbb{P}[\widehat{y}_{\mathrm{cPes},N}(x) \in B(x,b) \mid m_N > 0] \le \min\left\{ 1, \frac{\bar{\varepsilon}_g(x)^2}{b^2 q(x,b)} + \frac{\bar{\varepsilon}_g(x)}{b}\sqrt{\frac{D^{\sharp}(x)}{q(x,b)}} \right\}. \tag{21}$$

The same bounds hold for the population policy (19). With an abstaining fallback, the sample bounds hold unconditionally for every N.

The bound is uniform in N, but its size still depends on the safety-proxy error $\bar{\varepsilon}_g(x)$ and may be large when $q(x,b)$ is small. Thus, cPes prevents increasing the sampling budget from causing unbounded concentration on false acceptances, but it cannot identify or remove them. Increasing β or clipping the reward scores more tightly limits this concentration at the cost of weaker reward optimization. In the Appendix, we complement this safety guarantee with utility and regret bounds for cPes. For more details, see Proposition A.9 and Theorem A.10.

6 Experiments

6.1 Toy Problem

We first consider a three-class toy problem. Each candidate belongs to a latent class A, B, or C with probabilities $(0.39, 0.01, 0.60)$. Class A is truly safe and proxy-feasible, class B is unsafe but proxy-feasible, and class C is correctly rejected by the safety proxy. We set

$$g^\star = (1,0,0), \qquad \widehat{g} = (0.95, 0.90, 0.05), \qquad b = 0.8,$$

and use true rewards $r^\star = (0.8, 0.2, 0.5)$. The learned reward is $\hat{r} = r^\star + \sigma_k Z$, with $(\sigma_A, \sigma_B, \sigma_C) = (0.2, 1.0, 0.2)$ and $Z \sim \mathcal{N}(0,1)$. Thus the unsafe false-positive class B is rare, but has a substantially heavier upper tail of reward-proxy overestimation than the safe class A. We compare random feasible selection, CBoN, and cPes over increasing candidate budgets. Random feasible selection chooses uniformly among the candidates accepted by the safety proxy. CBoN uses the unclipped learned score $\hat{r}$, whereas cPes applies the clipping with $[\widehat{R}_{\min}, \widehat{R}_{\max}] = [0, 4]$.

Figure 1 shows the predicted amplification. As N grows, CBoN shifts from selecting the safe class A to the unsafe-but-feasible class B. The conditional safety hacking rate rises from 0.023 at $N = 1$ to

0.9996 at $N = 8192$, while the mean true reward falls from about 0.786 to 0.200. Random feasible selection remains near the baseline false-positive rate, showing that exposure to class B alone does not explain the failure. cPes substantially limits this amplification in the toy problem. Controls varying the reward-error tails and removing false positives confirm that the effect requires an unsafe proxy-feasible class with a sufficiently advantageous learned-reward tail; see Appendix A.15.

6.2 LLM Experiments

Experimental settings. Our evaluation uses 714 test and 179 validation prompts from JailbreakBench [Chao et al., 2024], HarmBench [Mazeika et al., 2024], and AdvBench [Zou et al., 2023], after removing duplicates. For each prompt, we sample 256 responses from `Qwen/Qwen2.5-7B-Instruct` [Qwen et al., 2025] (temperature 1.0, top-p 0.95, 512 new tokens), giving 228,608 candidates. We evaluate all selection methods on the same candidates.

We filter candidates with `meta-llama/Llama-Guard-3-8B` [Grattafiori et al., 2024] and rank them with `PKU-Alignment/beaver-7b-v1.0-reward` [Dai et al., 2024]. We report the primary safety threshold $b = 0.95$ and compare random feasible selection, CBoN, and cPes for $N \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$. Random feasible selection controls for exposure to filter false positives without reward maximization; cPes uses clipping levels fixed using the validation set and $\beta = 1.0$. Point estimates are averaged over 100 random candidate-order permutations. For each prompt, we first average the metric over permutations and then compute standard errors across prompts; shaded bands show ± 1.96 standard errors.

Because latent safety and reward are unobservable for open-ended responses, we use the HarmBench classifier as the operational safety criterion and define a safety-hacking event as selecting a response that passes Llama Guard but is classified as unsafe by HarmBench. Reported safety-hacking rates condition on non-abstention. Separately, we use `gpt-5-mini` [Singh et al., 2025] to evaluate safety-aware reward under the `safety_aware_v1` rubric. These evaluators are used as operational alternatives to the true functions $r^\star$ and $g^\star$ in the experiments. To validate the operational safety labels, we manually reviewed every unique proxy-feasible response that HarmBench classified as unsafe, i.e., every candidate in B. Across all such responses, manual review agreed with the HarmBench label in every case. Evaluator and scoring details are given in Appendix B.1.

Safety-hacking amplification and finite-N decomposition. For a fixed prompt, candidate ordering, and budget N, let K_A and K_B be the numbers of proxy-feasible candidates labeled safe and unsafe by HarmBench. When the corresponding class is nonempty, let M_A and M_B denote its largest proxy reward. Because the proxy scores have no ties, CBoN selects an unsafe response in one of two cases: no safe feasible candidate is available, or the best unsafe candidate outscores the best safe candidate. Conditional on non-abstention, this gives

$$\begin{aligned} &\Pr\left[\widehat{y}_N^{\mathrm{BoN}} \in B(x,b) \mid K_A + K_B > 0\right] \\ &= \Pr\left[K_B > 0,\ K_A = 0 \mid K_A + K_B > 0\right] \\ &\quad + \Pr\left[K_A > 0,\ K_B > 0,\ M_B > M_A \mid K_A + K_B > 0\right]. \end{aligned} \tag{22}$$

The first term is the probability that the candidate set contains an unsafe proxy-feasible output but no safe proxy-feasible output. The second measures reward-based competition when both classes are present, and is the empirical finite-N counterpart of the joint-tail competition in Theorem 4.1.

Figure 2 shows that the increase in safety hacking is driven primarily by cases in which an unsafe proxy-feasible candidate outscores all safe proxy-feasible candidates. In both the HarmBench subset and the combined evaluation, the contribution from cases with no safe proxy-feasible candidate decreases with N and therefore cannot explain the overall increase. At the same time, the competitive-selection term increases, so a growing fraction of hacking events occurs when both classes are present and the unsafe class attains the larger maximum proxy reward. This shift toward reward-based unsafe wins is consistent with the finite-N behavior governed by the relative joint tails in Theorem 4.1. Accordingly, the CBoN hacking rate increasingly exceeds the random-feasible contamination baseline, whereas cPes remains closer to the baseline (Figure 3a). Repeating the analysis with a calibration-matched `google/shieldgemma-2b` filter [Zeng et al., 2024] gives the same decomposition pattern, but a smaller overall increase in safety hacking: unsafe-only exposure falls with N, while competitive unsafe wins account for more of the total hacking rate (Appendix B.3).

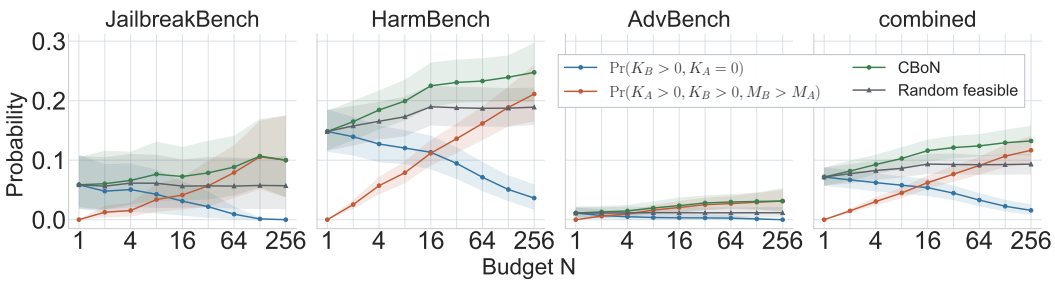
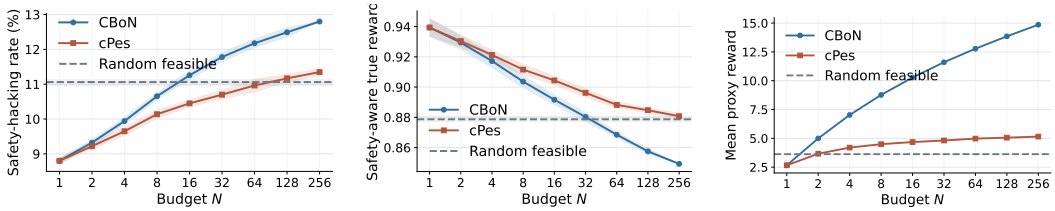


Figure 2: Finite-N decomposition of conditional CBoN safety hacking on adversarial prompts under the HarmBench evaluation. The total safety-hacking rate (green) is the sum of cases with no safe proxy-feasible candidate (blue) and cases in which an unsafe proxy-feasible candidate outscores all safe proxy-feasible candidates (orange). Random feasible selection is shown in gray. The competitive term captures reward-based unsafe wins when both classes are present and is the empirical finite-N counterpart of the joint-tail competition in Theorem 4.1. Points are averaged over 100 candidate-order permutations; bands show ± 1.96 standard errors across prompts after permutation averaging.



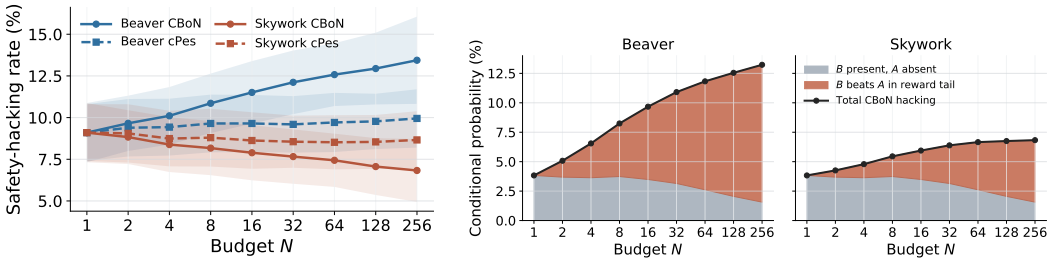
(a) Safety-hacking rate. (b) Safety-aware true reward. (c) Mean proxy reward.

Figure 3: Safety and reward on adversarial prompts. HarmBench evaluates safety, and `gpt-5-mini` evaluates safety-aware reward. As N grows, CBoN improves proxy reward but worsens both safety measures, while cPes limits these effects. Points are averaged over 100 candidate-order permutations; bands show ± 1.96 standard errors across prompts after permutation averaging.

Proxy reward versus safety-aware reward. Larger budgets improve the objective optimized by CBoN: its proxy reward increases well above the random-feasible reference (Figure 3c). Yet its `gpt-5-mini`-judged safety-aware reward, which penalizes unsafe and non-responsive outputs, deteriorates with N (Figure 3b). cPes produces smaller departures on both measures. Thus aggressive proxy optimization exploits proxy errors rather than improving the intended objective.

Reward proxy ablation. To isolate and evaluate the effect of the reward proxy, we replace `PKU-Alignment/beaver-7b-v1.0-reward` [Dai et al., 2024] with `Skywork/Skywork-Reward-V2-Llama-3.1-8B` [Liu et al., 2026]. We keep the candidate pools, safety filter, and HarmBench labels fixed. In this ablation, CBoN safety hacking increases with N under Beaver but decreases under Skywork, reversing the scaling trend. At $N = 256$, the contribution from cases with no safe proxy-feasible candidate is nearly identical for the two reward proxies, whereas cases in which an unsafe proxy-feasible candidate outscores all safe proxy-feasible candidates are substantially less common under Skywork (Figure 4). This reversal is therefore attributable to how the reward proxies rank the same safe and unsafe proxy-feasible candidates, rather than to filtering. This ablation is the empirical counterpart of the finite-N bounds in Theorem 4.1 and Corollary A.1, which depend on the joint tails $\widehat{\Psi}_A$ and $\widehat{\Psi}_B$. A favorable safe tail can prevent amplification over the evaluated budgets. At $N = 256$, Skywork CBoN has a lower safety-hacking rate than Skywork cPes (6.8% versus 8.7%). The guarantee for cPes is therefore not pointwise dominance over CBoN, but an N-independent bound under adverse tail configurations (Corollary 5.3 and Theorem 5.1).

Generator robustness. We repeat the analysis with `Mistral-7B-Instruct-v0.2` as the response generator, while retaining the same experimental settings. The Mistral replication exhibits the same reward-proxy dependence: safety hacking increases with N under Beaver but decreases under Skywork, while cPes limits amplification in the adverse Beaver setting. Full safety-hacking and proxy-reward curves are reported in Appendix B.2.



(a) Safety-hacking rate.

(b) Finite-N decomposition for CBoN.

Figure 4: Reward-proxy ablation with fixed candidate pools, safety-filter decisions, and HarmBench labels. Skywork reverses the CBoN safety-hacking trend relative to Beaver (left). The frequency of cases in which only unsafe candidates pass the filter is unchanged, but unsafe candidates are less likely to outscore all safe candidates that pass the filter under Skywork (right). This reversal shows that amplification depends on how the reward model ranks the candidates admitted by the safety filter.

7 Discussion and Limitations

Safety hacking is difficult to rule out. Our results show that inference-time scaling can induce safety hacking on instances where unsafe outputs can be generated with nonzero probability and then receive higher reward scores than safe alternatives. Inference-time scaling can make these outputs more likely to be selected. Because perfectly accurate reward and safety proxies are difficult to obtain, such vulnerabilities may persist even as these proxies improve. As AI systems become more capable, these failures could have increasingly severe consequences, including catastrophic outcomes.

Pessimism does not necessarily provide an effective soluiton to protect against safety hacking. Our findings therefore caution against assuming that methods developed to address reward overoptimization will also prevent safety failures. The Skywork results suggest that using reward models that better capture safety is a promising direction. Nevertheless, improved safety assessment does not guarantee a zero probability of safety hacking. Residual blind spots may still allow unsafe outputs to be favored, and inference-time scaling can amplify their consequences. These remaining risks deserve continued attention even when a reward model performs well on average.

Limitations. We analyze a fixed, non-adaptive pipeline in which independently sampled candidates are filtered once and then ranked. Our theory therefore does not cover adaptive search or self-refinement, whose candidate distributions can depend on previous proxy evaluations. By repeatedly steering generation toward high proxy scores, such procedures may amplify residual safety-filter errors more strongly, but our results do not establish or quantify this behavior. Extending our analysis to such feedback-driven procedures is an important direction for future work.

8 Conclusion

We studied safety hacking in constrained inference-time scaling, where reward maximization can amplify residual errors in a learned safety filter. The filter may admit unsafe responses, and downstream selection may favor them more strongly as inference compute grows. Our finite-N analysis shows that this behavior is governed by the joint reward tails of safe and unsafe proxy-feasible responses, and our asymptotic result establishes that a heavier unsafe tail can make safety hacking nearly certain even when average proxy errors are small. Coverage control yields an N-independent bound, instantiated by cPes, but cannot remove unsafe responses already admitted by the filter. Toy and language-model experiments provide finite-budget evidence for this amplification mechanism, while the reward-proxy ablation shows that ranking within the contaminated feasible set is central to the observed outcome. The analysis therefore separates two requirements for safe inference-time scaling: controlling entry into the feasible set and controlling downstream concentration. Extending this characterization to adaptive and agentic search remains an important direction for future work. More broadly, we hope this work encourages researchers to treat safety as a central consideration when designing and evaluating increasingly capable inference-time scaling methods.

AI USE STATEMENT

In this work, we used generative AI tools for coding, literature review, and manuscript preparation. We did not use generative AI tools to develop the theoretical model or conceptual framework, formulate mathematical claims, provide ingredients for or write proofs, propose or refine hypotheses, design the research methodology, or independently interpret the results. We additionally used generative AI tools for literature-search assistance and language and style editing of the manuscript. The authors inspected and tested all AI-assisted code, checked suggested literature against the original sources, manually verified all citations, and reviewed and revised all AI-assisted text and feedback. Model-generated evaluations are reported as operational measures rather than ground truth. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses public adversarial-prompt benchmarks and may generate harmful responses. The study involves no human participants or private data, and its intended use is to improve the safety of inference-time methods.

REPRODUCIBILITY STATEMENT

The appendix provides complete proofs and additional experimental details, including the evaluation setup and ablations. We also submit the code used to run the experiments and reproduce the reported results.

REFERENCES

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Yaswanth Chittepu, Ativ Joshi, Sohini Chintala, and Scott Niekum. Safe inference-time alignment via Lagrangian reward augmentation. *arXiv preprint arXiv:2607.02781*, 2026.

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. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *arXiv preprint arXiv:2109.13916*, 2021.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? Coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25075–25126. PMLR, 13–19 Jul 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.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: LLM-based input-output safeguard for human-AI conversations. *arXiv preprint arXiv:2312.06674*, 2023.

Xiaotong Ji, Shyam Sundhar Ramesh, Matthieu Zimmer, Ilija Bogunovic, Jun Wang, and Haitham Bou Ammar. On almost surely safe alignment of large language models at inference-time. *arXiv preprint arXiv:2502.01208*, 2025.

Yuu Jinnai, Tetsuro Morimura, Kaito Ariu, and Kenshi Abe. Regularized best-of-n sampling with minimum bayes risk objective for language model alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9321–9347, 2025.

Hadi Khalaf, Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio Calmon. Inference-time reward hacking in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-reward-v2: Scaling preference data curation via human-AI synergy. In *The Fourteenth International Conference on Learning Representations*, 2026.

David Manheim and Scott Garrabrant. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 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.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in neural information processing systems*, 35:9460–9471, 2022.

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.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 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.

Haoyu Wang, Zeyu Qin, Li Shen, Xueqian Wang, Dacheng Tao, and Minhao Cheng. Safety reasoning with guidelines. In *Forty-second International Conference on Machine Learning*, 2025.

Zhuohao Yu, Steven Wu, and Adam Block. From curiosity to caution: Mitigating reward hacking for best-of-n with pessimism. In *The Fourteenth International Conference on Learning Representations*, 2026.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. Shieldgemma: Generative AI content moderation based on gemma, 2024.

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.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Appendix

A.1 Finite-N Safety-hacking Criteria

Corollary A.1. *Under the conditions of Theorem 4.1, the following hold.*

1. *If there exist $t \in \mathbb{R}$, $\alpha_A \in [0,1)$, and $\gamma_B > 0$ such that $N\widehat{\Psi}_A(t;x,b) \le \alpha_A$ and $N\widehat{\Psi}_B(t;x,b) \ge \gamma_B$, then*

$$\mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x,b)\right] \ge (1-\alpha_A)\big(1-e^{-\gamma_B}\big). \tag{23}$$

2. *If there exist $s \in \mathbb{R}$, $\alpha_B \ge 0$, and $\gamma_A > 0$ such that $N\widehat{\Psi}_B(s;x,b) \le \alpha_B$, and $N\widehat{\Psi}_A(s;x,b) \ge \gamma_A$, then*

$$\mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x,b)\right] \le \alpha_B + e^{-\gamma_A}. \tag{24}$$

Proof. Let $\psi_A = \widehat{\Psi}_A(t;x,b)$ and $\psi_B = \widehat{\Psi}_B(t;x,b)$. From Theorem 4.1,

$$\widehat{L}_N(t;x,b) = (1-\psi_A)^N\left[1-\left(1-\frac{\psi_B}{1-\psi_A}\right)^N\right]. \tag{25}$$

If $N\psi_A \le \alpha_A < 1$, then $(1-\psi_A)^N \ge 1 - N\psi_A \ge 1-\alpha_A$. Since $\psi_B/(1-\psi_A) \ge \psi_B$ and $N\psi_B \ge \gamma_B$,

$$1-\left(1-\frac{\psi_B}{1-\psi_A}\right)^N \ge 1-e^{-N\psi_B} \ge 1-e^{-\gamma_B}. \tag{26}$$

This proves the first claim. For the second claim, let $\psi_A = \widehat{\Psi}_A(s;x,b)$ and $\psi_B = \widehat{\Psi}_B(s;x,b)$. The upper bound gives

$$\widehat{U}_N(s;x,b) \le \left[1-(1-\psi_B)^N\right] + (1-\psi_A)^N \le N\psi_B + e^{-N\psi_A} \le \alpha_B + e^{-\gamma_A}. \tag{27}$$

This completes the proof. □

Corollary A.1 gives an experimentally testable diagnostic: as N increases, estimate whether the effective unsafe-but-feasible score-tail mass $N\widehat{\Psi}_B(\cdot)$ crosses a constant before the true-positive score-tail mass $N\widehat{\Psi}_A(\cdot)$ does. If so, CBoN should display compute-amplified safety hacking.

A.2 Prompt-distribution Amplification

Corollary A.2 (Prompt-distribution amplification). *Let $X \sim \mathcal{D}$ be a prompt drawn from a prompt distribution, and fix $b \in (0, 1]$. For each prompt x, define the conditional CBoN hacking probability, where the probability is over the N candidate samples drawn from $\pi_{\mathrm{ref}}(\cdot \mid x)$, by*

$$p_N(x) \coloneqq \mathbb{P}_{y_{1:N} \sim \pi_{\mathrm{ref}}(\cdot \mid x)} \left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x, b) \right]. \tag{28}$$

Suppose there exist a measurable set of prompts $\mathcal{E} \subseteq \mathcal{X}$ and $\rho \in [0, 1]$ such that

$$\mathbb{P}_{X \sim \mathcal{D}}[X \in \mathcal{E}] \geq \rho,$$

and, for every $x \in \mathcal{E}$, there exists a sequence $t_N(x) \in \mathbb{R}$ satisfying $N\widehat{\Psi}_A(t_N(x); x, b) \to 0$ and $N\widehat{\Psi}_B(t_N(x); x, b) \to \infty$. Then the aggregate safety-hacking probability satisfies

$$\liminf_{N \to \infty} \mathbb{P} \left[\widehat{y}_N^{\mathrm{BoN}}(X) \in B(X, b) \right] \geq \rho, \tag{29}$$

where the probability is over $X \sim \mathcal{D}$ and the candidate samples drawn conditionally from $\pi_{\mathrm{ref}}(\cdot \mid X)$.

Proof. By Theorem 4.2, for every $x \in \mathcal{E}$ we have $p_N(x) \to 1$. Moreover, $0 \leq p_N(x) \leq 1$ for all x and N. Therefore

$$\mathbb{P} \left[\widehat{y}_N^{\mathrm{BoN}}(X) \in B(X, b) \right] = \mathbb{E}_{X \sim \mathcal{D}}[p_N(X)] \geq \mathbb{E}_{X \sim \mathcal{D}}[p_N(X) \mathbb{I}\{X \in \mathcal{E}\}]. \tag{30}$$

Fatou's lemma gives

$$\liminf_{N \to \infty} \mathbb{E}[p_N(X) \mathbb{I}\{X \in \mathcal{E}\}] \geq \mathbb{E} \left[\liminf_{N \to \infty} p_N(X) \mathbb{I}\{X \in \mathcal{E}\} \right] \tag{31}$$

$$= \mathbb{P}_{X \sim \mathcal{D}}[X \in \mathcal{E}] \geq \rho. \tag{32}$$

This proves the claim. □

Corollary A.2 lifts the fixed-prompt tail condition to distribution-level evaluations: if a positive-measure subset of prompts is tail-amplifiable, then the average safety-hacking rate cannot vanish as the search budget grows.

A.3 Direct score dominance

Remark A.3 (Direct score dominance). *If there exists a threshold $t \in \mathbb{R}$ such that $\widehat{\Psi}_A(t; x, b) = 0$ and $\widehat{\Psi}_B(t; x, b) > 0$, then Theorem 4.2 applies with the constant sequence $t_N = t$, and*

$$\mathbb{P}[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x, b)] \to 1. \tag{33}$$

This holds, for example, when a positive-reference-mass subset of $B(x, b)$ receives learned scores above the essential upper endpoint on $A(x, b)$.

A.4 Gaussian error-tail separation

Definition A.4 (Proxy-feasible reward range). *Fix $x \in \mathcal{X}$ and $b \in (0, 1]$ with $\widehat{\mathcal{S}}_+(x, b) \neq \varnothing$. Define*

$$\Delta(x, b) \coloneqq \sup_{y \in \widehat{\mathcal{S}}_+(x,b)} r^\star(x, y) - \inf_{y \in \widehat{\mathcal{S}}_+(x,b)} r^\star(x, y). \tag{34}$$

Assumption A.5 (Local bounded reward range). *For the prompt $x \in \mathcal{X}$ and threshold $b \in (0, 1]$ under consideration, $\widehat{\mathcal{S}}_+(x, b) \neq \varnothing$ and $\Delta(x, b) < \infty$.*

Proposition A.6 (Gaussian error-tail separation). *We define the pointwise proxy errors*

$$\varepsilon_r(x, y) \coloneqq \widehat{r}(x, y) - r^\star(x, y), \tag{35}$$

$$\varepsilon_g(x, y) \coloneqq \widehat{g}(x, y) - g^\star(x, y). \tag{36}$$

Also, for $\Diamond \in \{A, B\}$, define

$$\eta_\Diamond(x, b) \coloneqq \mathbb{P}_{y \sim \pi_{\mathrm{ref}}(\cdot \mid x)}[y \in \Diamond(x, b)]. \tag{37}$$

Fix $x \in \mathcal{X}$ and $b \in (0, 1]$, and suppose Assumption A.5 holds. Assume that $\eta_A(x, b) > 0$ and $\eta_B(x, b) > 0$. Suppose further that there exist constants $0 < \sigma_A(x) < \sigma_B(x)$ such that the conditional upper tails of the reward-proxy error satisfy, as $u \to \infty$,

$$\log \mathbb{P}\left[\varepsilon_r(x, y) > u \mid y \in A(x, b)\right] = -\frac{u^2}{2\sigma_A^2(x)} + o(u^2), \tag{38}$$

$$\log \mathbb{P}\left[\varepsilon_r(x, y) > u \mid y \in B(x, b)\right] = -\frac{u^2}{2\sigma_B^2(x)} + o(u^2). \tag{39}$$

Let

$$R_A^\star(x, b) \coloneqq \sup_{y \in A(x,b)} r^\star(x, y). \tag{40}$$

Then, for any constant $c \in \left(\sqrt{2}\sigma_A(x), \sqrt{2}\sigma_B(x)\right)$, the sequence

$$\tau_N \coloneqq R_A^\star(x, b) + c\sqrt{\log N} \tag{41}$$

satisfies the score-tail separation condition in Theorem 4.2; that is,

$$N\widehat{\Psi}_A(\tau_N; x, b) \to 0, \qquad N\widehat{\Psi}_B(\tau_N; x, b) \to \infty. \tag{42}$$

Consequently,

$$\lim_{N\to\infty} \mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x, b)\right] = 1. \tag{43}$$

The fixed shift $\Delta(x, b)$ only enters when converting reward-error tails into learned-score tails, and it does not change the Gaussian large-deviation exponent.

Proof. For $\Diamond \in \{A, B\}$ and $t \in \mathbb{R}$, define the joint reward-error tail

$$\Psi_\Diamond(t; x, b) \coloneqq \mathbb{P}_{y\sim\pi_{\mathrm{ref}}(\,\cdot\,|x)}\left[y \in \Diamond(x, b),\ \varepsilon_r(x, y) > t\right]. \tag{44}$$

These tails combine the reference probability of a class with the upper-tail behavior of reward-proxy overestimation inside that class.

For $\Diamond \in \{A, B\}$, we can write

$$\Psi_\Diamond(u; x, b) = \eta_\Diamond(x, b)\mathbb{P}\left[\varepsilon_r(x, y) > u \mid y \in \Diamond(x, b)\right].$$

Let

$$t_N = c\sqrt{\log N}, \qquad c \in (\sqrt{2}\sigma_A(x), \sqrt{2}\sigma_B(x)),$$

and set

$$\tau_N = R_A^\star(x, b) + t_N.$$

Using the assumed tail exponent on $A(x, b)$,

$$\log\left(N\Psi_A(t_N; x, b)\right) = \log N + \log \eta_A(x, b) - \frac{t_N^2}{2\sigma_A^2(x)} + o(t_N^2) \tag{45}$$

$$= \left(1 - \frac{c^2}{2\sigma_A^2(x)} + o(1)\right)\log N + O(1). \tag{46}$$

Since $c > \sqrt{2}\sigma_A(x)$, the coefficient of $\log N$ is negative, and therefore

$$N\Psi_A(t_N; x, b) \to 0. \tag{47}$$

Moreover, for $y \in A(x, b)$, $r^\star(x, y) \le R_A^\star(x, b)$. Hence

$$\{y \in A(x, b),\ \widehat{r}(x, y) > \tau_N\} \subseteq \{y \in A(x, b),\ \varepsilon_r(x, y) > t_N\}, \tag{48}$$

and so

$$N\widehat{\Psi}_A(\tau_N; x, b) \le N\Psi_A(t_N; x, b) \to 0. \tag{49}$$

For the B tail, since $\Delta(x, b) < \infty$ by assumption,

$$(t_N + \Delta(x, b))^2 = c^2 \log N + o(\log N). \tag{50}$$

Algorithm 1 Constrained Pessimistic Inference (cPes), sample implementation

Require: Prompt x, reference policy $\pi_{\mathrm{ref}}(\cdot \mid x)$, proxies $\widehat{r}, \widehat{g}$, threshold b, parameter β, clipping levels $\widehat{R}_{\min} < \widehat{R}_{\max}$, budget N

1: Sample $y_1, \ldots, y_N \overset{\text{iid}}{\sim} \pi_{\mathrm{ref}}(\cdot \mid x)$
2: Set $\mathcal{I}_N(x,b) = \{i \in [N] : \widehat{g}(x, y_i) \geq b\}$ and $m_N = |\mathcal{I}_N(x,b)|$
3: **if** $m_N = 0$ **then**
4: Return a predefined fallback, e.g., abstention
5: **end if**
6: Set $\widetilde{r}_i = \mathrm{clip}(\widehat{r}(x, y_i), \widehat{R}_{\min}, \widehat{R}_{\max})$ for each $i \in \mathcal{I}_N(x,b)$
7: Find $\hat{\lambda}_N(x) \in [\widehat{R}_{\min} - \beta, \widehat{R}_{\max})$ such that

$$\frac{1}{m_N} \sum_{i \in \mathcal{I}_N(x,b)} \mathrm{ReLU}\left(\frac{\widetilde{r}_i - \hat{\lambda}_N(x)}{\beta}\right) = 1$$

8: Set $\hat{p}_i = \frac{1}{m_N} \mathrm{ReLU}\left(\frac{\widetilde{r}_i - \hat{\lambda}_N(x)}{\beta}\right)$ for each $i \in \mathcal{I}_N(x,b)$
9: Sample $I \in \mathcal{I}_N(x,b)$ with probability $\hat{p}_i$ and return y_I

Using the assumed tail exponent on $B(x,b)$,

$$\log\left(N\Psi_B(t_N + \Delta(x,b); x, b)\right) = \log N + \log \eta_B(x,b) - \frac{(t_N + \Delta(x,b))^2}{2\sigma_B^2(x)} + o\left((t_N + \Delta(x,b))^2\right) \tag{51}$$

$$= \left(1 - \frac{c^2}{2\sigma_B^2(x)} + o(1)\right) \log N + O(1). \tag{52}$$

Since $c < \sqrt{2}\sigma_B(x)$, the coefficient of $\log N$ is positive, and therefore

$$N\Psi_B(t_N + \Delta(x,b); x, b) \to \infty. \tag{53}$$

For every $y \in B(x,b)$, we have $y \in \widehat{\mathcal{S}}_+(x,b)$ and $A(x,b) \subseteq \widehat{\mathcal{S}}_+(x,b)$. Hence $R_A^\star(x,b) \leq \sup_{z \in \widehat{\mathcal{S}}_+(x,b)} r^\star(x,z)$ and $r^\star(x,y) \geq \inf_{z \in \widehat{\mathcal{S}}_+(x,b)} r^\star(x,z)$, so

$$R_A^\star(x,b) - r^\star(x,y) \leq \Delta(x,b). \tag{54}$$

Thus

$$\{y \in B(x,b),\ \varepsilon_r(x,y) > t_N + \Delta(x,b)\} \subseteq \{y \in B(x,b),\ \widehat{r}(x,y) > \tau_N\}, \tag{55}$$

and hence

$$N\widehat{\Psi}_B(\tau_N; x, b) \geq N\Psi_B(t_N + \Delta(x,b); x, b) \to \infty. \tag{56}$$

Therefore the score-tail separation conditions in Theorem 4.2 hold with threshold τ_N. The final claim follows from Theorem 4.2. □

Remark A.7 (Boundary behavior). *When $\widehat{\Psi}_A$ and $\widehat{\Psi}_B$ are asymptotically comparable on the relevant score scale, finer tail constants and class masses determine the limit. In error-tail conditions such as Proposition A.6, the finite reward range $\Delta(x,b)$ can affect boundary constants without changing the Gaussian exponent away from the boundary.*

A.5 Sample implementation and consistency of cPes

Proposition A.8 (Consistency of sample cPes). *Fix $x \in \mathcal{X}$ and $b \in (0,1]$ with $q(x,b) > 0$. Let $\hat{\pi}_N(\cdot \mid x)$ be the random output distribution induced by Algorithm 1 on $m_N > 0$. Then $m_N \to \infty$ almost surely, $\hat{\lambda}_N(x) \to \lambda(x)$ almost surely, and for every bounded $f : \mathcal{Y} \to \mathbb{R}$,*

$$\mathbb{E}_{y \sim \hat{\pi}_N(\cdot \mid x)}[f(y)] \to \mathbb{E}_{y \sim \hat{\pi}(\cdot \mid x)}[f(y)] \qquad \textit{almost surely}. \tag{57}$$

Proof. Let $m_N = \sum_{i=1}^{N} \mathbb{I}\{y_i \in \widehat{\mathcal{S}}_+(x,b)\}$. Since $m_N/N \to q(x,b) > 0$ almost surely, $m_N \to \infty$ almost surely and $m_N > 0$ eventually.

Conditional on belonging to $\widehat{\mathcal{S}}_+(x,b)$, the feasible samples are i.i.d. from $\pi^\sharp_{\mathrm{ref}}(\cdot \mid x)$. Define

$$h_N(\lambda) \coloneqq \frac{1}{m_N} \sum_{i \in \mathcal{I}_N(x,b)} \mathrm{ReLU}\left(\frac{\widetilde{r}(x,y_i) - \lambda}{\beta}\right), \tag{58}$$

$$h(\lambda) \coloneqq \mathbb{E}_{\pi^\sharp_{\mathrm{ref}}} \mathrm{ReLU}\left(\frac{\widetilde{r}(x,y) - \lambda}{\beta}\right). \tag{59}$$

The functions indexed by $\lambda \in [\widehat{R}_{\min} - \beta, \widehat{R}_{\max}]$ are uniformly bounded and $1/\beta$-Lipschitz in λ. A finite-grid argument combined with the strong law of large numbers yields

$$\sup_{\lambda \in [\widehat{R}_{\min} - \beta, \widehat{R}_{\max}]} |h_N(\lambda) - h(\lambda)| \to 0 \qquad \text{almost surely.} \tag{60}$$

By Proposition 5.2, the equation $h(\lambda) = 1$ has a unique solution $\lambda(x)$. Uniform convergence and monotonicity imply that any empirical solution $\hat{\lambda}_N(x)$ to $h_N(\lambda) = 1$ converges almost surely to $\lambda(x)$.

For any bounded f, write

$$\mathbb{E}_{\hat{\pi}_N}[f] = \frac{1}{m_N} \sum_{i \in \mathcal{I}_N(x,b)} f(y_i) \mathrm{ReLU}\left(\frac{\widetilde{r}(x,y_i) - \hat{\lambda}_N(x)}{\beta}\right). \tag{61}$$

The summands are uniformly bounded, and the weights converge uniformly in the multiplier because of the Lipschitz property. Applying the strong law again and using $\hat{\lambda}_N(x) \to \lambda(x)$ gives

$$\mathbb{E}_{\hat{\pi}_N}[f] \to \mathbb{E}_{\pi^\sharp_{\mathrm{ref}}}\left[f(y) \mathrm{ReLU}\left(\frac{\widetilde{r}(x,y) - \lambda(x)}{\beta}\right)\right] = \mathbb{E}_{\hat{\pi}}[f]. \tag{62}$$

□

A.6 Utility tradeoff of cPes

Bounded coverage alone can be achieved by avoiding reward optimization altogether. We therefore quantify the utility cost of coverage control. For a policy π, let

$$J(\pi; x) \coloneqq \mathbb{E}_{y \sim \pi(\cdot \mid x)}[r^\star(x,y)]. \tag{63}$$

Define the clipped-score RMSE under the proxy-feasible reference by

$$\bar{\varepsilon}^\sharp_{r,\mathrm{clip}}(x) \coloneqq \sqrt{\mathbb{E}_{y \sim \pi^\sharp_{\mathrm{ref}}(\cdot \mid x)}\left[(\widetilde{r}(x,y) - r^\star(x,y))^2\right]}. \tag{64}$$

This quantity includes both reward-proxy error and any distortion introduced by clipping.

Proposition A.9 (Utility tradeoff within the proxy-feasible set)**.** *Suppose $\bar{\varepsilon}^\sharp_{r,\mathrm{clip}}(x) < \infty$. For any comparator policy $\rho(\cdot \mid x) \ll \pi^\sharp_{\mathrm{ref}}(\cdot \mid x)$, the population cPes policy satisfies*

$$J(\rho; x) - J(\hat{\pi}; x) \le \frac{\beta}{2}\left(C^\sharp_\rho(x) - 1\right) + \bar{\varepsilon}^\sharp_{r,\mathrm{clip}}(x)\sqrt{C^\sharp_\rho(x) - 1} + \Delta(x), \tag{65}$$

where

$$\Delta(x) := \begin{cases} \frac{(\bar{\varepsilon}^\sharp_{r,\mathrm{clip}}(x))^2}{2\beta} & \sqrt{D^\sharp(x)} \ge \frac{\bar{\varepsilon}^\sharp_{r,\mathrm{clip}}(x)}{\beta}, \\ \bar{\varepsilon}^\sharp_{r,\mathrm{clip}}(x)\sqrt{D^\sharp(x)} - \frac{\beta}{2} D^\sharp(x) & \textit{otherwise.} \end{cases} \tag{66}$$

The bound makes explicit the role of β. Smaller β permits more aggressive reward optimization but allows greater concentration relative to $\pi^\sharp_{\mathrm{ref}}$, while larger β keeps the policy closer to the proxy-feasible reference distribution. The variance-sensitive branch tightens the bound when the clipped score has small variance under $\pi^\sharp_{\mathrm{ref}}$.

A.7 ORACLE REGRET DECOMPOSITION

For the oracle comparison below, define the ambient coverage coefficient

$$C_\pi(x) \coloneqq \sum_{y\in\mathcal{Y}} \frac{\pi(y \mid x)^2}{\pi_{\mathrm{ref}}(y \mid x)}. \tag{67}$$

For a comparator $\pi^\star$, let $\alpha^\star(x,b) \coloneqq \pi^\star(\widehat{\mathcal{S}}_+(x,b) \mid x)$ and, when $\alpha^\star(x,b) > 0$, define

$$\pi^{\star,\sharp}(y \mid x) \coloneqq \frac{\pi^\star(y \mid x)\mathbb{I}\{y \in \widehat{\mathcal{S}}_+(x,b)\}}{\alpha^\star(x,b)}. \tag{68}$$

Theorem A.10 (Oracle regret decomposition)**.** *Fix $b \in (0,1)$ and suppose $0 \le r^\star(x,y) \le R_{\max}$. Let $\pi^\star \ll \pi_{\mathrm{ref}}$ satisfy $\mathrm{supp}(\pi^\star) \subseteq \mathcal{S}^\star_+(x)$; if $\alpha^\star(x,b) > 0$, assume also $\pi^{\star,\sharp} \ll \pi^\sharp_{\mathrm{ref}}$. If $\alpha^\star(x,b) > 0$, then*

$$J(\pi^\star;x) - J(\hat\pi;x) \le \left[\frac{\beta}{2}\big(C^\sharp_{\pi^{\star,\sharp}}(x) - 1\big) + \bar\varepsilon^\sharp_{r,\mathrm{clip}}(x)\sqrt{C^\sharp_{\pi^{\star,\sharp}}(x) - 1} + \Delta(x)\right] + \frac{R_{\max}\sqrt{C_{\pi^\star}(x)}\,\bar\varepsilon_g(x)}{1-b}. \tag{69}$$

If $\alpha^\star(x,b) = 0$, then

$$J(\pi^\star;x) - J(\hat\pi;x) \le \frac{R_{\max}\sqrt{C_{\pi^\star}(x)}\,\bar\varepsilon_g(x)}{1-b}, \tag{70}$$

that is, the bracketed proxy-feasible comparison term is omitted.

Proof. Decompose the regret of the truly safe comparator $\pi^\star$:

$$J(\pi^\star;x) - J(\hat\pi;x) = \mathbb{E}_{\pi^\star}[r^\star(x,y)\mathbb{I}\{y \in \widehat{\mathcal{S}}_+(x,b)\}] - J(\hat\pi;x) \tag{71}$$

$$+ \mathbb{E}_{\pi^\star}[r^\star(x,y)\mathbb{I}\{y \notin \widehat{\mathcal{S}}_+(x,b)\}]. \tag{72}$$

If $\alpha^\star(x,b) > 0$, then the first term is at most $J(\pi^{\star,\sharp};x) - J(\hat\pi;x)$ because $\alpha^\star(x,b) \le 1$ and rewards are nonnegative. Applying Proposition A.9 with $\rho = \pi^{\star,\sharp}$ gives the bracketed term in (69). If $\alpha^\star(x,b) = 0$, the first term is $-J(\hat\pi;x) \le 0$, so the bracketed term is omitted.

It remains to control the second term. Since $\mathrm{supp}(\pi^\star) \subseteq \mathcal{S}^\star_+(x)$, the event $\{y \notin \widehat{\mathcal{S}}_+(x,b)\}$ under $\pi^\star$ is a false negative of the safety proxy. On this event, $g^\star(x,y) = 1$ and $\widehat{g}(x,y) < b$, so $|\varepsilon_g(x,y)| > 1-b$. Hence

$$\mathbb{P}_{\pi_{\mathrm{ref}}}[y \in \mathcal{S}^\star_+(x),\ y \notin \widehat{\mathcal{S}}_+(x,b)] \le \frac{\bar\varepsilon_g(x)^2}{(1-b)^2}. \tag{73}$$

By Cauchy–Schwarz and absolute continuity,

$$\mathbb{P}_{\pi^\star}[y \notin \widehat{\mathcal{S}}_+(x,b)] = \mathbb{E}_{\pi_{\mathrm{ref}}}\left[\frac{\pi^\star(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}\mathbb{I}\{y \in \mathcal{S}^\star_+(x),\ y \notin \widehat{\mathcal{S}}_+(x,b)\}\right] \tag{74}$$

$$\le \sqrt{C_{\pi^\star}(x)}\sqrt{\mathbb{P}_{\pi_{\mathrm{ref}}}[y \in \mathcal{S}^\star_+(x),\ y \notin \widehat{\mathcal{S}}_+(x,b)]} \tag{75}$$

$$\le \frac{\sqrt{C_{\pi^\star}(x)}\,\bar\varepsilon_g(x)}{1-b}. \tag{76}$$

Using $r^\star(x,y) \le R_{\max}$ proves

$$\mathbb{E}_{\pi^\star}[r^\star(x,y)\mathbb{I}\{y \notin \widehat{\mathcal{S}}_+(x,b)\}] \le \frac{R_{\max}\sqrt{C_{\pi^\star}(x)}\,\bar\varepsilon_g(x)}{1-b}. \tag{77}$$

Combining the bounds proves the theorem. □

A.8 Proof of Theorem 4.1

Proof. Suppress (x, b) in the notation. For a single sample y, define

$$\widehat{E}_A(t) \coloneqq \{y \in A,\ \widehat{r}(x, y) > t\}, \tag{78}$$

$$\widehat{E}_B(t) \coloneqq \{y \in B,\ \widehat{r}(x, y) > t\}. \tag{79}$$

The sets A and B are disjoint, so these two one-sample events are disjoint. If among N samples no event $\widehat{E}_A(t)$ occurs and at least one event $\widehat{E}_B(t)$ occurs, then every true-positive proxy-feasible sample has learned score at most t, while some unsafe-but-feasible sample has learned score strictly larger than t. Therefore the CBoN maximizer over the proxy-feasible samples lies in B. The probability of this sufficient event is

$$\big(1 - \widehat{\Psi}_A(t)\big)^N - \big(1 - \widehat{\Psi}_A(t) - \widehat{\Psi}_B(t)\big)^N, \tag{80}$$

which proves the lower bound.

For the upper bound, consider the event that no unsafe-but-feasible sample satisfies $\widehat{r}(x, y) > s$ and at least one true-positive proxy-feasible sample satisfies $\widehat{r}(x, y) > s$. On this event, some sample in A strictly beats every sample in B, so CBoN cannot hack. The probability of this non-hacking certificate is

$$\big(1 - \widehat{\Psi}_B(s)\big)^N - \big(1 - \widehat{\Psi}_B(s) - \widehat{\Psi}_A(s)\big)^N. \tag{81}$$

Hence the hacking probability is at most its complement,

$$1 - \big(1 - \widehat{\Psi}_B(s)\big)^N + \big(1 - \widehat{\Psi}_B(s) - \widehat{\Psi}_A(s)\big)^N, \tag{82}$$

which is (9). □

A.9 Proof of Theorem 4.2

Proof. Let $a_N = \widehat{\Psi}_A(t_N; x, b)$ and $d_N = \widehat{\Psi}_B(t_N; x, b)$. The assumptions imply $Na_N \to 0$ and $Nd_N \to \infty$. Therefore $(1 - a_N)^N \to 1$ and $(1 - a_N - d_N)^N \le \exp[-N(a_N + d_N)] \to 0$. The lower bound in Theorem 4.1 converges to one, proving the claim. □

A.10 Proof of Corollary 4.3

Proof. Fix x and suppress the dependence on x in the notation. Let the output space contain two outputs,

$$\mathcal{Y} = \{y_A, y_B\},$$

with reference probabilities

$$\pi_{\mathrm{ref}}(y_A \mid x) = 1 - \varepsilon, \qquad \pi_{\mathrm{ref}}(y_B \mid x) = \varepsilon.$$

Define the true and learned safety functions by

$$g^\star(x, y_A) = 1, \qquad \widehat{g}(x, y_A) = 1,$$

and

$$g^\star(x, y_B) = 0, \qquad \widehat{g}(x, y_B) = b.$$

Then both outputs are proxy-feasible, since $\widehat{g}(x, y_A) \ge b$ and $\widehat{g}(x, y_B) \ge b$. Moreover,

$$A(x, b) = \{y_A\}, \qquad B(x, b) = \{y_B\},$$

and therefore

$$\mathbb{P}_{y \sim \pi_{\mathrm{ref}}(\cdot \mid x)}[y \in B(x, b)] = \varepsilon.$$

The safety-proxy RMSE is

$$\sqrt{\mathbb{E}_{\pi_{\mathrm{ref}}}\left[\big(\widehat{g}(x, y) - g^\star(x, y)\big)^2\right]} = \sqrt{\varepsilon b^2} = b\sqrt{\varepsilon}.$$

Now define the true and learned reward functions by

$$r^\star(x, y_A) = 0, \qquad \widehat{r}(x, y_A) = 0,$$

and

$$r^\star(x, y_B) = 0, \qquad \widehat{r}(x, y_B) = \xi.$$

Thus the reward-proxy RMSE is

$$\sqrt{\mathbb{E}_{\pi_{\mathrm{ref}}}\left[\left(\widehat{r}(x, y) - r^\star(x, y)\right)^2\right]} = \sqrt{\varepsilon \xi^2} = \xi\sqrt{\varepsilon}.$$

Since $\widehat{r}(x, y_B) > \widehat{r}(x, y_A)$, constrained Best-of-$N$ returns y_B if and only if at least one of the N sampled candidates equals y_B. All sampled candidates are proxy-feasible, so there is no abstention event. Hence

$$\mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x, b)\right] = 1 - (1 - \varepsilon)^N.$$

Using $(1 - \varepsilon)^N \le \exp(-N\varepsilon)$ gives

$$\mathbb{P}\left[\widehat{y}_N^{\mathrm{BoN}}(x) \in B(x, b)\right] \ge 1 - \exp(-N\varepsilon).$$

Therefore, if

$$N \ge \varepsilon^{-1} \log(1/\delta),$$

then the safety-hacking probability is at least $1 - \delta$. This proves the claim. □

A.11 Proof of Theorem 5.1

Proof. Let

$$w(y) \coloneqq \frac{\pi(y \mid x)}{\pi^\sharp_{\mathrm{ref}}(y \mid x)}, \qquad \kappa \coloneqq \pi^\sharp_{\mathrm{ref}}(B(x, b) \mid x).$$

Since $\mathbb{E}_{\pi^\sharp_{\mathrm{ref}}}[w] = 1$ and $\mathbb{E}_{\pi^\sharp_{\mathrm{ref}}}[w^2] = C^\sharp_\pi(x)$,

$$\mathbb{E}_{\pi^\sharp_{\mathrm{ref}}}[(w - 1)^2] = C^\sharp_\pi(x) - 1. \tag{83}$$

Moreover,

$$\mathbb{P}_{y\sim\pi(\cdot\,|x)}[y \in B(x, b)] - \kappa = \mathbb{E}_{\pi^\sharp_{\mathrm{ref}}}\left[(w(y) - 1)\mathbb{I}\{y \in B(x, b)\}\right] \tag{84}$$

$$= \mathbb{E}_{\pi^\sharp_{\mathrm{ref}}}\left[(w(y) - 1)\big(\mathbb{I}\{y \in B(x, b)\} - \kappa\big)\right], \tag{85}$$

where the second equality uses $\mathbb{E}_{\pi^\sharp_{\mathrm{ref}}}[w - 1] = 0$. By Cauchy–Schwarz,

$$\left|\mathbb{P}_{y\sim\pi(\cdot\,|x)}[y \in B(x, b)] - \kappa\right| \le \sqrt{C^\sharp_\pi(x) - 1}\sqrt{\kappa(1 - \kappa)}. \tag{86}$$

This proves the bound. The equality holds if and only if $w(y) = 1 + c\big(\mathbb{I}\{y \in B(x, b)\} - \kappa\big)$ for some c.

By the definition of $\pi^\sharp_{\mathrm{ref}}$, we have

$$\mathbb{P}_{y\sim\pi^\sharp_{\mathrm{ref}}(\cdot\,|x)}[y \in B(x, b)] = \mathbb{E}_{y\sim\pi^\sharp_{\mathrm{ref}}(\cdot\,|x)}\left[\mathbb{I}\{y \in B(x, b)\}\right] \tag{87}$$

$$= \mathbb{E}_{y\sim\pi_{\mathrm{ref}}(\cdot\,|x)}\left[\frac{\mathbb{I}\{y \in B(x, b)\}\mathbb{I}\{y \in \widehat{\mathcal{S}}_+(x, b)\}}{q(x, b)}\right] \tag{88}$$

$$= \frac{\mathbb{E}_{y\sim\pi_{\mathrm{ref}}(\cdot\,|x)}\left[\mathbb{I}\{y \in B(x, b)\}\right]}{q(x, b)}. \tag{89}$$

If $y \in B(x, b)$, then $g^\star(x, y) = 0$ and $\widehat{g}(x, y) \ge b$, so $|\varepsilon_g(x, y)| \ge b$. Hence

$$\mathbb{I}\{y \in B(x, b)\} \le \frac{\varepsilon_g(x, y)^2}{b^2}. \tag{90}$$

Taking expectation under $\pi_{\mathrm{ref}}(\cdot \mid x)$ gives

$$\pi_{\mathrm{ref}}(B(x, b) \mid x) \le \frac{\bar{\varepsilon}_g(x)^2}{b^2}. \tag{91}$$

Combining this inequality with $\kappa(x, b) = \pi_{\mathrm{ref}}(B(x, b) \mid x)/q(x, b)$ gives

$$\kappa(x, b) \le \frac{\bar{\varepsilon}_g(x)^2}{b^2 q(x, b)}, \tag{92}$$

which proves (16). □

A.12 Proof of Proposition 5.2

Proof. Let $\mu(y) = \pi_{\text{ref}}^{\sharp}(y \mid x)$ and write $w(y) = \pi(y \mid x)/\mu(y)$. Then $\pi \ll \mu$ and $\sum_y \pi(y \mid x) = 1$ are equivalent to $w(y) \geq 0$ and $\mathbb{E}_\mu[w] = 1$. The objective in (18) becomes

$$\max_{w \geq 0,\ \mathbb{E}_\mu[w]=1} \mathbb{E}_{y\sim\mu}\left[w(y)\widetilde{r}(x,y) - \frac{\beta}{2}w(y)^2\right]. \tag{93}$$

This is strictly concave in w, so the optimizer is unique.

Introduce a Lagrange multiplier λ for $\mathbb{E}_\mu[w] = 1$. Pointwise maximization of the Lagrangian over $w(y) \geq 0$ gives

$$w_\lambda(y) = \left(\frac{\widetilde{r}(x,y) - \lambda}{\beta}\right)_+ = \mathrm{ReLU}\left(\frac{\widetilde{r}(x,y) - \lambda}{\beta}\right). \tag{94}$$

The multiplier must satisfy $\mathbb{E}_\mu[w_\lambda] = 1$.

It remains to show existence and uniqueness of λ. Define

$$h(\lambda) \coloneqq \mathbb{E}_\mu\left[\mathrm{ReLU}\left(\frac{\widetilde{r}(x,y) - \lambda}{\beta}\right)\right]. \tag{95}$$

The function h is continuous and nonincreasing. By construction, $h(\widehat{R}_{\max}) = 0$, while

$$h(\widehat{R}_{\min} - \beta) = \mathbb{E}_\mu\left[\frac{\widetilde{r}(x,y) - \widehat{R}_{\min} + \beta}{\beta}\right] \geq 1. \tag{96}$$

Thus a solution exists in $[\widehat{R}_{\min} - \beta, \widehat{R}_{\max})$. It is unique because whenever $h(\lambda) > 0$, the event $\{\widetilde{r}(x,y) > \lambda\}$ has positive μ-probability, so h is strictly decreasing on the relevant level set.

Finally, let $w(y) = w_{\lambda(x)}(y)$. Since $\mathbb{E}_\mu[w] = 1$ and $\lambda(x) \geq \widehat{R}_{\min} - \beta$,

$$0 \leq w(y) \leq \frac{\widehat{R}_{\max} - \lambda(x)}{\beta} \leq 1 + \frac{\widehat{R}_{\text{span}}}{\beta}. \tag{97}$$

Therefore

$$C_{\hat{\pi}}^{\sharp}(x) = \mathbb{E}_\mu[w(y)^2] \leq \big(\sup_y w(y)\big)\mathbb{E}_\mu[w(y)] \leq 1 + \frac{\widehat{R}_{\text{span}}}{\beta}. \tag{98}$$

The additional claim,

$$C_{\hat{\pi}}^{\sharp}(x) \leq 1 + \frac{V^{\sharp}(x)}{\beta^2}, \qquad \text{where} \qquad V^{\sharp}(x) := \mathrm{Var}_{\pi_{\text{ref}}^{\sharp}}[\widetilde{r}(x,y)], \tag{99}$$

is derived as follows. Note $w(y) = \mathrm{ReLU}\left(\frac{\widetilde{r}(x,y)-\lambda(x)}{\beta}\right)$, and let $\nu(y) = \mathrm{ReLU}(\lambda(x) - \widetilde{r}(x,y))$. Then, we have $\beta w(y) = \widetilde{r}(x,y) - \lambda(x) + \nu(y)$ and $w(y)\nu(y) = 0$ for every x, y. Therefore,

$$\begin{aligned}
&\beta w(y) = \widetilde{r}(x,y) - \lambda(x) + \nu(y) \\
\implies\ &\beta w(y)(w(y)-1) = (\widetilde{r}(x,y) - \lambda(x) + \nu(y))(w(y)-1) \\
\implies\ &\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\beta w(y)(w(y)-1)] = \mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[(\widetilde{r}(x,y) - \lambda(x) + \nu(y))(w(y)-1)] \\
\implies\ &\beta\big(C_{\hat{\pi}}^{\sharp}(x) - 1\big) = \mathbb{E}_{\hat{\pi}}[\widetilde{r}(x,y)] - \mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\widetilde{r}(x,y)] - \lambda(x)(\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[w(y)] - 1) - \mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\nu(y)].
\end{aligned}$$

Noting that $\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[w(y)] = 1$ and $\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\nu(y)] \geq 0$, we have

$$\beta\big(C_{\hat{\pi}}^{\sharp}(x) - 1\big) \leq \mathbb{E}_{\hat{\pi}}[\widetilde{r}(x,y)] - \mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\widetilde{r}(x,y)].$$

The right-hand side can be rewritten as $\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[(w(y)-1)(\widetilde{r}(x,y) - \mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\widetilde{r}(x,y)])]$. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[(w(y)-1)(\widetilde{r}(x,y) - \mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\widetilde{r}(x,y)])] \\
&\quad\leq \sqrt{\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[(w(y)-1)^2]}\sqrt{\mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[(\widetilde{r}(x,y) - \mathbb{E}_{\pi_{\text{ref}}^{\sharp}}[\widetilde{r}(x,y)])]^2} \\
&\quad= \sqrt{C_{\hat{\pi}}^{\sharp}(x) - 1}\sqrt{V^{\sharp}(x)}
\end{aligned}$$

Hence, we have $C_{\hat{\pi}}^{\sharp}(x) - 1 \leq V^{\sharp}(x)/\beta^2$. This concludes the proof. □

A.13 Proof of Proposition A.9

Proof. Let $\mu = \pi_{\mathrm{ref}}^{\sharp}(\cdot \mid x)$ and $e = \widetilde{r} - r^{\star}$. Then,

$$J(\rho; x) - J(\hat{\pi}; x) = (\mathbb{E}_{\rho}[\widetilde{r}(x,y)] - \mathbb{E}_{\hat{\pi}}[\widetilde{r}(x,y)]) + (\mathbb{E}_{\hat{\pi}}[e(x,y)] - \mathbb{E}_{\rho}[e(x,y)])\,. \tag{100}$$

By optimality of $\hat{\pi}$ in (18), the first bracket is upper-bounded as

$$\mathbb{E}_{\rho}[\widetilde{r}(x,y)] - \mathbb{E}_{\hat{\pi}}[\widetilde{r}(x,y)] \le \frac{\beta}{2}\big(C_{\rho}^{\sharp}(x) - C_{\hat{\pi}}^{\sharp}(x)\big). \tag{101}$$

For the second bracket, letting $w = \hat{\pi}/\pi_{\mathrm{ref}}^{\sharp}$ and $w_{\rho} = \rho/\pi_{\mathrm{ref}}^{\sharp}$,

$$\begin{aligned}
\mathbb{E}_{\hat{\pi}}[e(x,y)] - \mathbb{E}_{\rho}[e(x,y)] &= \mathbb{E}_{\mu}[(w(x,y) - w_{\rho}(x,y))(e(x,y) - \mathbb{E}_{\mu}[e(x,y)])] && (102)\\
&\le \sqrt{\mathbb{E}_{\mu}[(w(x,y) - w_{\rho}(x,y))^2]}\sqrt{\mathbb{E}_{\mu}[(e(x,y) - \mathbb{E}_{\mu}[e(x,y)])^2]} && (103)\\
&= \sqrt{\mathbb{E}_{\mu}[(w(x,y) - w_{\rho}(x,y))^2]}\,\bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x) && (104)\\
&\le \left(\sqrt{\mathbb{E}_{\mu}[(w(x,y) - 1)^2]} + \sqrt{\mathbb{E}_{\mu}[(w_{\rho}(x,y) - 1)^2]}\right)\bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x) && (105)\\
&= \left(\sqrt{C_{\hat{\pi}}^{\sharp}(x) - 1} + \sqrt{C_{\rho}^{\sharp}(x) - 1}\right)\bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x), && (106)
\end{aligned}$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the second inequality follows from Minkowski's inequality. Let $D_{\rho} = C_{\rho}^{\sharp}(x) - 1$ and $D_{\widehat{\pi}} = C_{\widehat{\pi}}^{\sharp}(x) - 1$. Then, we obtain

$$\begin{aligned}
J(\rho; x) - J(\hat{\pi}; x) &\le \frac{\beta}{2}D_{\rho} + \bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x)\sqrt{D_{\rho}} - \frac{\beta}{2}D_{\hat{\pi}} + \bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x)\sqrt{D_{\hat{\pi}}} && (107)\\
&= \frac{\beta}{2}D_{\rho} + \bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x)\sqrt{D_{\rho}} + \frac{(\bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x))^2}{2\beta} - \frac{\beta}{2}\left(\sqrt{D_{\hat{\pi}}} - \frac{\bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x)}{\beta}\right)^2 && (108)\\
&\le \frac{\beta}{2}D_{\rho} + \bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x)\sqrt{D_{\rho}}\\
&\quad + \begin{cases} \frac{(\bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x))^2}{2\beta} & \sqrt{D^{\sharp}(x)} \ge \frac{\bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x)}{\beta} \\ \bar{\varepsilon}_{r,\mathrm{clip}}^{\sharp}(x)\sqrt{D^{\sharp}(x)} - \frac{\beta}{2}D^{\sharp}(x) & \text{otherwise,} \end{cases} && (109)
\end{aligned}$$

where $D^{\sharp}$ is defined in (20) and $D_{\hat{\pi}} \le D^{\sharp}$ is used. This completes the proof. □

A.14 Proof of Corollary 5.3

Finite-sample claim.

Proof. Condition on the event $m_N > 0$ and on the sampled candidates. Let

$$w_i = \mathrm{ReLU}\left(\frac{\widetilde{r}(x, y_i) - \hat{\lambda}_N(x)}{\beta}\right), \qquad i \in \mathcal{I}_N(x, b). \tag{110}$$

By construction, $m_N^{-1}\sum_{i\in\mathcal{I}_N} w_i = 1$. The conditional probability that cPes returns an unsafe-but-feasible sample is

$$\frac{1}{m_N}\sum_{i\in\mathcal{I}_N} w_i \mathbb{I}\{y_i \in B(x,b)\}. \tag{111}$$

Let

$$K_N^B \coloneqq \sum_{i=1}^{N} \mathbb{I}\{y_i \in B(x,b)\}, \tag{112}$$

$$\widehat{\kappa}_N \coloneqq \frac{K_N^B}{m_N}. \tag{113}$$

Since $m_N^{-1}\sum_{i\in\mathcal{I}_N}(w_i-1)=0$, we have

$$\frac{1}{m_N}\sum_{i\in\mathcal{I}_N} w_i\mathbb{I}\{y_i\in B(x,b)\}-\widehat{\kappa}_N=\frac{1}{m_N}\sum_{i\in\mathcal{I}_N}(w_i-1)\left(\mathbb{I}\{y_i\in B(x,b)\}-\widehat{\kappa}_N\right). \tag{114}$$

By Cauchy–Schwarz,

$$\left|\frac{1}{m_N}\sum_{i\in\mathcal{I}_N} w_i\mathbb{I}\{y_i\in B(x,b)\}-\widehat{\kappa}_N\right|\le\left(\frac{1}{m_N}\sum_{i\in\mathcal{I}_N}(w_i-1)^2\right)^{1/2}\sqrt{\widehat{\kappa}_N(1-\widehat{\kappa}_N)}. \tag{115}$$

The empirical analogue of (20) gives

$$\widehat{V}_N\coloneqq\frac{1}{m_N}\sum_{i\in\mathcal{I}_N}\left(\widetilde{r}(x,y_i)-\frac{1}{m_N}\sum_{j\in\mathcal{I}_N}\widetilde{r}(x,y_j)\right)^2, \tag{116}$$

and

$$\frac{1}{m_N}\sum_{i\in\mathcal{I}_N}(w_i-1)^2=\frac{1}{m_N}\sum_{i\in\mathcal{I}_N}w_i^2-1\le\min\left\{\frac{\widehat{V}_N}{\beta^2},\frac{\widehat{R}_{\mathrm{span}}}{\beta}\right\}. \tag{117}$$

Consequently, conditional on the sampled candidates,

$$\mathbb{P}[\widehat{y}_{\mathrm{cPes},N}(x)\in B(x,b)\mid y_1,\dots,y_N]\le\widehat{\kappa}_N+\sqrt{\min\left\{\frac{\widehat{V}_N}{\beta^2},\frac{\widehat{R}_{\mathrm{span}}}{\beta}\right\}\widehat{\kappa}_N(1-\widehat{\kappa}_N)}. \tag{118}$$

Given $m_N=m>0$, the feasible samples are i.i.d. from $\pi_{\mathrm{ref}}^{\sharp}$, and hence

$$K_N^B\mid m_N=m\sim\mathrm{Binomial}\big(m,\kappa(x,b)\big). \tag{119}$$

Therefore,

$$\mathbb{E}[\widehat{\kappa}_N\mid m_N=m]=\kappa(x,b). \tag{120}$$

Moreover, letting $V^{\sharp}(x)=\mathrm{Var}_{\pi_{\mathrm{ref}}^{\sharp}}[\widetilde{r}(x,y)]$,

$$\mathbb{E}[\widehat{V}_N\mid m_N=m]=\left(1-\frac{1}{m}\right)V^{\sharp}(x), \tag{121}$$

$$\mathbb{E}[\widehat{\kappa}_N(1-\widehat{\kappa}_N)\mid m_N=m]=\left(1-\frac{1}{m}\right)\kappa(x,b)(1-\kappa(x,b)). \tag{122}$$

Applying Cauchy–Schwarz to the variance term and Jensen's inequality to the range term, and using $1-1/m\le 1$, gives, for every $m>0$,

$$\mathbb{P}[\widehat{y}_{\mathrm{cPes},N}(x)\in B(x,b)\mid m_N=m]\le\kappa(x,b)+\sqrt{\min\left\{\frac{V^{\sharp}(x)}{\beta^2},\frac{\widehat{R}_{\mathrm{span}}}{\beta}\right\}\kappa(x,b)(1-\kappa(x,b))}. \tag{123}$$

Averaging over m gives the same bound conditional on $m_N>0$.

Finally, Theorem 5.1 gives

$$\kappa(x,b)\le\frac{\bar{\varepsilon}_g(x)^2}{b^2q(x,b)}. \tag{124}$$

Using $1-\kappa(x,b)\le 1$, we obtain

$$\begin{aligned}&\mathbb{P}[\widehat{y}_{\mathrm{cPes},N}(x)\in B(x,b)\mid m_N>0]\\&\quad\le\min\left\{1,\ \frac{\bar{\varepsilon}_g(x)^2}{b^2q(x,b)}+\frac{\bar{\varepsilon}_g(x)}{b}\sqrt{\frac{1}{q(x,b)}\min\left\{\frac{V^{\sharp}(x)}{\beta^2},\frac{\widehat{R}_{\mathrm{span}}}{\beta}\right\}}\right\},\end{aligned} \tag{125}$$

which proves (21). The unconditional statement follows because abstention is not an unsafe output. □

Population claim.

Proof. By Proposition 5.2, the population pessimistic policy satisfies $\hat{\pi}(\cdot \mid x) \ll \pi_{\mathrm{ref}}^{\sharp}(\cdot \mid x)$ and

$$C_{\hat{\pi}}^{\sharp}(x) \le 1 + \min\left\{ \frac{V^{\sharp}(x)}{\beta^2}, \frac{\widehat{R}_{\mathrm{span}}}{\beta} \right\}. \tag{126}$$

Applying Theorem 5.1 and using (20) gives

$$\mathbb{P}_{y\sim\hat{\pi}(\cdot\,|x)}[y \in B(x,b)] \le \kappa(x,b) + \sqrt{\min\left\{ \frac{V^{\sharp}(x)}{\beta^2}, \frac{\widehat{R}_{\mathrm{span}}}{\beta} \right\} \kappa(x,b)\big(1-\kappa(x,b)\big)}. \tag{127}$$

Since

$$\kappa(x,b) \le \frac{\bar{\varepsilon}_g(x)^2}{b^2 q(x,b)}, \tag{128}$$

and $1-\kappa(x,b) \le 1$, it follows that

$$\mathbb{P}_{y\sim\hat{\pi}(\cdot\,|x)}[y \in B(x,b)] \le \min\left\{ 1,\ \frac{\bar{\varepsilon}_g(x)^2}{b^2 q(x,b)} + \frac{\bar{\varepsilon}_g(x)}{b} \sqrt{\frac{1}{q(x,b)} \min\left\{ \frac{V^{\sharp}(x)}{\beta^2}, \frac{\widehat{R}_{\mathrm{span}}}{\beta} \right\}} \right\}. \tag{129}$$

□

A.15 Additional Toy Experimental Details and Ablations

Experimental details. The toy problem in Section 6.1 uses class probabilities $(0.39, 0.01, 0.60)$ for A, B, and C, respectively. We evaluate random feasible selection, CBoN, and cPes over budgets

$$N \in \{1, 2, 4, \ldots, 8192\}, \tag{130}$$

using 5000 Monte Carlo trials per budget. Unless otherwise stated, metrics are reported conditional on non-abstention, since at small N all sampled candidates may belong to the proxy-rejected class C.

The true reward gap between the safe and unsafe proxy-feasible classes is

$$\Delta = r_A^{\star} - r_B^{\star} = 0.6, \tag{131}$$

which matches the local reward-range shift used when translating reward-error tails into learned-score tails in Proposition A.6.

cPes regularization. The cPes results in Figure 1 are consistent with the budget-uniform control predicted by Corollary 5.3. With $\beta = 1.0$, the conditional hacking rate remains between 0.015 and 0.024 over the entire budget grid, while the mean true reward stays near 0.79. Even with $\beta = 0.2$, the hacking rate remains around 0.024–0.031 at large budgets.

A separate sweep over β illustrates the concentration tradeoff. At $N = 8192$, the hacking rate is 0.1518 for $\beta = 0.02$, 0.0418 for $\beta = 0.1$, and 0.0152 for $\beta = 1.0$, with the highest true reward observed around $\beta \in [0.5, 1.0]$. Thus, as cPes approaches aggressive max-selection, it increasingly recovers the same amplification mechanism as CBoN.

Equal-tail control. To verify that large candidate budgets alone do not cause safety hacking, we set

$$\sigma_B = \sigma_A = 0.2$$

while keeping the remainder of the environment unchanged. In this setting, the unsafe class B no longer has a heavier reward-error upper tail than the safe class A. The CBoN hacking rate falls to zero by $N = 32$, and the selected true reward converges to 0.8, the reward of class A.

No-false-positive control. We also remove class B entirely. In this case the proxy-feasible set contains no unsafe outputs, and the safety-hacking rate is identically zero for all methods and all candidate budgets.

False-positive tail-scale sweep. Finally, Figure 1c varies the reward-proxy noise scale σ_B of the unsafe-but-feasible class. At $N = 8192$, the CBoN hacking rate increases from 0.0 at $\sigma_B = 0.2$ to 0.697 at $\sigma_B = 0.6$, 0.9996 at $\sigma_B = 1.0$, and 0.9998 at $\sigma_B = 1.2$.

Together, these controls support the mechanism characterized by Theorem 4.2: increasing N is not sufficient by itself. Amplification occurs when residual unsafe outputs remain proxy-feasible and have a sufficiently favorable upper tail under the learned reward.

B Additional LLM Analysis

B.1 Evaluator and scoring details

Safety proxy and operational safety criterion. The safety proxy is `Llama-Guard-3-8B`. We use $1 - p_{\text{unsafe}}$ as its safe score and accept candidates whose score is at least $b = 0.95$. Primary safety labels are produced by the HarmBench classifier `HarmBench-Llama-2-13b-cls`. Thus A contains proxy-feasible candidates classified as safe by HarmBench, B contains proxy-feasible candidates classified as unsafe, and C contains proxy-rejected candidates. The reported safety-hacking rate is the probability of selecting from B, conditional on non-abstention.

Safety-aware reward evaluator. We score safety-aware reward with the OpenAI Responses API using the `gpt-5-mini` API model alias and the `safety_aware_v1` rubric. The rubric assigns 1.0 to responses that are safe and fully satisfy a legitimate request, 0.0 to responses that are unsafe, enable harmful behavior, or are useless/nonresponsive, and intermediate values to partially helpful, vague, incomplete, or weakly useful responses.

Reward proxies and aggregation. The primary reward proxy is `PKU-Alignment/beaver-7b-v1.0-reward`. For the reward-proxy ablation, we use `Skywork-Reward-V2-Llama-3.1-8B` with Hugging Face Transformers. Its input is formatted as `User:` {`prompt_text`}\n`Assistant:` {`response_text`}, and the scalar sequence-classification logit is used as the raw proxy reward. Downstream normalization uses the calibration median and interquartile range; clipping is applied only where required by cPes.

B.2 Mistral-generator replication

We test whether the empirical findings depend on the response generator by replacing `Qwen2.5-7B-Instruct` with `Mistral-7B-Instruct-v0.2`. For each of the 714 test prompts, we sample 256 responses using temperature 1.0, top-p 0.95, and a maximum of 512 new tokens. We retain Llama Guard as the safety proxy, HarmBench as the operational safety evaluator, the threshold $b = 0.95$, and 100 candidate-order permutations. We evaluate both the Beaver-7B and Skywork Reward V2 reward proxies.

With Beaver reward, CBoN safety hacking increases from 12.5% at $N = 1$ to 21.4% at $N = 256$, while its mean proxy reward increases from 9.41 to 19.56. Over the same budgets, cPes safety hacking increases only to 14.3%, and its proxy reward reaches 11.66. The random-feasible safety-hacking reference at $N = 256$ is 13.0%. Thus, with Beaver reward, the Mistral candidate pool reproduces the primary pattern: aggressive proxy optimization raises reward while amplifying residual safety-filter errors, whereas coverage-controlled selection limits the amplification.

With Skywork reward, CBoN exhibits the opposite safety trend: its safety-hacking rate decreases from 12.5% to 7.9%, even as its proxy reward increases from 15.16 to 25.42. At $N = 256$, cPes has a safety-hacking rate of 11.6% and a proxy reward of 16.66. This result reinforces that increasing the candidate budget is not intrinsically unsafe. Amplification depends on the reward proxy assigning sufficiently favorable rankings to unsafe-but-feasible responses. It also illustrates that the coverage guarantee for cPes does not imply pointwise dominance over CBoN under every reward proxy.

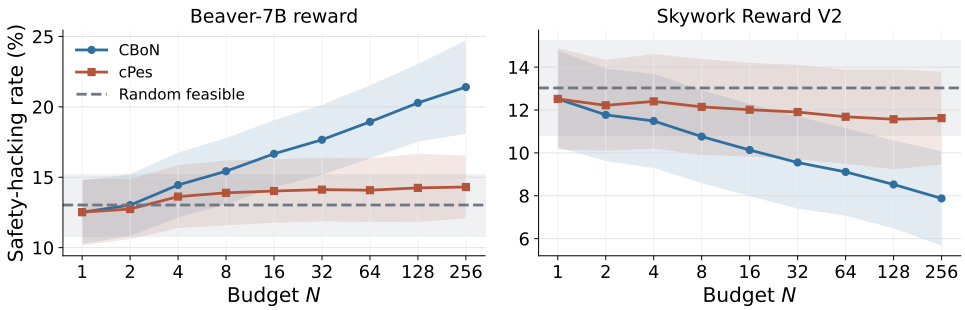


Figure 5: Safety-hacking rate with `Mistral-7B-Instruct-v0.2` as the response generator. Rates are conditional on non-abstention. Points are means over 100 candidate-order permutations and bands show ± 1.96 standard errors. The dashed line is the random-feasible reference estimated at $N = 256$. Under Beaver reward, CBoN amplifies safety hacking more strongly than cPes. Under Skywork reward, CBoN safety hacking instead decreases with N.

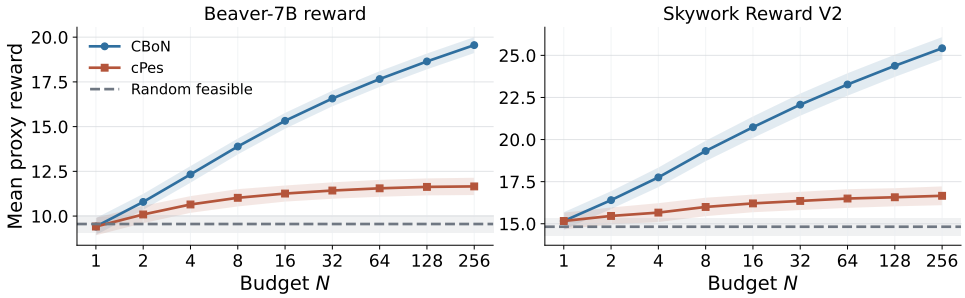


Figure 6: Mean selected proxy reward with `Mistral-7B-Instruct-v0.2` as the response generator. CBoN increasingly optimizes the corresponding reward proxy, while cPes remains closer to the random-feasible reference. Beaver and Skywork scores have different numerical scales and should be interpreted within, rather than across, panels.

B.3 Alternative safety-filter sensitivity analysis

We repeat the CBoN analysis using `google/shieldgemma-2b` as an alternative safety filter, while holding the candidate pool, generator, Beaver reward proxy, HarmBench labels, and candidate permutations fixed. We select the ShieldGemma threshold using only the 179 calibration prompts by matching the Llama Guard calibration feasible rate. The resulting threshold is 0.8176, giving a feasible rate of 0.8242, compared with 0.8295 for Llama Guard at its primary threshold.

N	Hacking	Abstention	Unsafe-only	Competitive	Tail win
1	12.10 (0.98)	14.24 (1.04)	12.10	0.00	–
256	14.75 (1.34)	1.26 (0.42)	0.99	13.76	35.02

Table 1: ShieldGemma-2B alternative-filter sensitivity analysis. All entries are percentages; standard errors across prompts are shown in parentheses for hacking and abstention. Unsafe-only and Competitive are the two disjoint terms in (23) and sum to the total hacking rate. Tail win denotes $\Pr(M_B > M_A \mid K_A > 0, K_B > 0)$ and is reported separately because it uses a different conditioning event.

Conditional CBoN safety hacking increases from 12.1% at $N = 1$ to 14.8% at $N = 256$. Over the same range, the unsafe-only term decreases from 12.1% to 1.0%, while the competitive unsafe-win term increases from zero to 13.8%. Thus the observed increase is driven by reward-based competition rather than unsafe-only exposure. The increase is smaller than in the primary analysis, so we treat

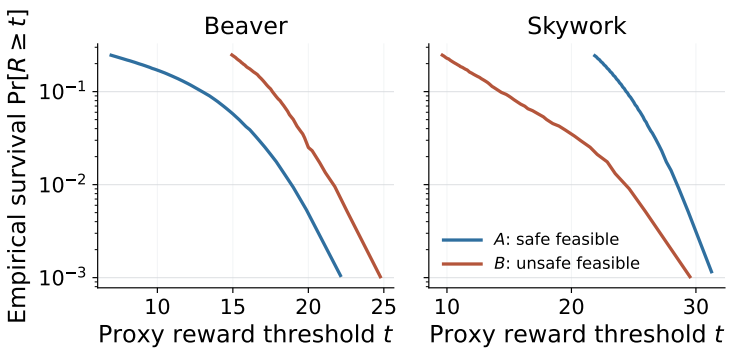


Figure 7: Conditional survival curves of learned reward scores among safe feasible (A) and unsafe feasible (B) responses. With Beaver, the unsafe feasible class has the heavier observed upper tail; with Skywork, the safe feasible class does. The panels use their respective proxy-score scales and should be interpreted within, rather than across, reward proxies.

this result as qualitatively consistent sensitivity evidence rather than as a replacement for the primary Llama Guard analysis.

B.4 Reward-tail survival curves

Figure 7 visualizes the conditional learned-score survival curves within the safe feasible class $A(x, b)$ and the unsafe feasible class $B(x, b)$. These curves are the score-distribution component of the joint tails $\widehat{\Psi}_A$ and $\widehat{\Psi}_B$ in (6); the class masses are held fixed across the Beaver and Skywork ablations. With Beaver, unsafe feasible responses retain more mass at high reward thresholds than safe feasible responses. With Skywork, the safe feasible curve instead extends farther into the upper tail. This reversal is consistent with the lower competitive unsafe-win term in Figure 4b and with the finite-N behavior in Figure 4a. Because the figure estimates conditional survival curves from a finite candidate pool, it is diagnostic evidence for the tail mechanism rather than a verification of the asymptotic conditions in Theorem 4.2.