EFFICIENT BEST-OF-N POLICY EVALUATION FOR INFERENCE-TIME ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-of-N (BoN) is a common inference-time alignment method that selects the highest-scoring response among N samples from a reference model. Evaluating BoN policies from logged data is challenging under sample-only access because standard off-policy estimators require density ratios that depend on unavailable response likelihoods. In this paper, we propose a sample-only framework for evaluating and selecting BoN policies without access to these likelihoods. We show that the order-statistic structure of BoN allows the required density ratios to be expressed through score-rank probabilities that are estimable from samples alone. We then develop a doubly robust estimator of the BoN policy value (BoN-DR) that efficiently reuses a shared auxiliary sample pool across candidate budgets. We establish valid asymptotic inference even under reward estimator misspecification and prove the efficiency of our BoN-DR estimator. Since larger budgets can amplify errors in the score function and lead to reward overoptimization, we derive two selection rules: (i) maximizing the estimated policy value and (ii) maximizing a lower confidence bound on the improvement over the reference policy, which accounts for estimation uncertainty and provides a no-harm guarantee. Across synthetic experiments and GSM8K with multiple reference and reward models, our framework accurately estimates BoN policy values and selects effective sampling budgets.

1 INTRODUCTION

In *inference-time alignment*, the aim is to improve the behavior of a given generative model using sampled outputs and an auxiliary reward or preference signal, but *without* updating or accessing the model parameters (Huang et al., 2025; Khalaf et al., 2025). This is relevant in practice, where one often interacts with a model through a restricted interface and cannot modify the underlying parameters. For example, many large companies want to adapt frontier LLMs with API-only access to their own preferences and needs (Kosmas et al., 2025), such as a car manufacturer using collected user feedback to improve a third-party in-car assistant.

A widely used approach is *Best-of-N* (BoN) sampling (Nakano et al., 2021; Beirami et al., 2025), which has two steps: (1) for a prompt x, the reference model (e.g., the LLM) generates N candidate responses; and (2) a score function s (e.g., an externally pre-trained reward model or an LLM judge) selects the highest-scoring response (Nakano et al., 2021; Beirami et al., 2025). Importantly, BoN requires **only** samples and **no** modification of the reference model. However, deploying BoN in practice is subject to two challenges:

Challenge (1)***:*** *Evaluating BoN policies under sample-only access.* Practitioners need to know whether and, in particular, *how much* a BoN policy actually improves the true downstream reward (and not only the score) over the reference model. From logged responses and rewards, this is an *off-policy* evaluation (OPE) problem. Yet, standard OPE estimators (e.g., importance-weighted and doubly robust estimators) are **not** applicable. The reason is that these estimators require the density ratio between the target policy and the policy that generated the logged data (Dudik et al., 2011). Under sample-only access, however, the probabilities assigned to complete responses, i.e., the response likelihoods, are unavailable. $\Rightarrow$ *We therefore need an estimator that evaluates BoN policies* ***without*** *evaluating the response likelihoods.*

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.

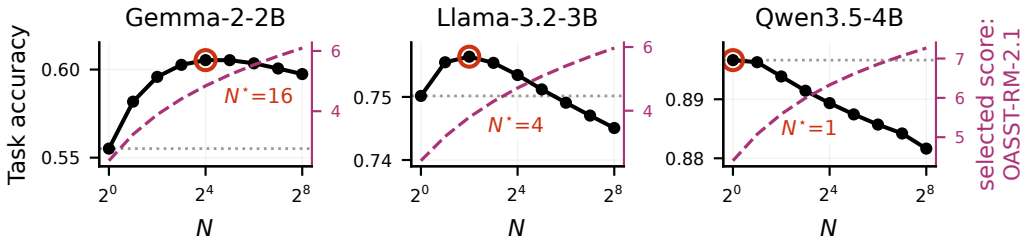


Figure 1: **The score function does not reveal the optimal inference-time budget.** Best-of-N on GSM8K (Cobbe et al., 2021) for three reference models using OASST-RM-2.1 as the score model. **Black solid:** task accuracy computed from 512 stored responses for each of the 1,319 held-out test prompts. **Magenta dashed:** mean selected-response score (right axis). **Grey dotted:** reference-model accuracy ($N = 1$). ⊙ Red circles mark the accuracy-optimal budget $N^\star$. ⇒ **Takeaway:** The proxy score increases with N, but accuracy peaks and then falls, showing that the score function alone cannot determine the best sampling budget.

Challenge ② *: Selecting the sampling budget N.* The sampling budget N controls how strongly BoN optimizes the score: a small N limits the potential benefit, while a large N concentrates the policy on low-probability, high-scoring responses. When the score function is an imperfect proxy for the downstream reward, this can amplify its errors and reduce downstream performance. This issue is often known as *reward overoptimization* (Gao et al., 2023; Coste et al., 2024). Figure 1 illustrates the issue on GSM8K: the score function of the selected response increases with N, while the downstream accuracy can peak and then decline. The score function alone therefore does not reveal which N should be deployed; instead, this requires estimating the policy value of each candidate budget. However, even unbiased value estimates are noisy in finite samples, especially for large N, where the BoN policy concentrates on responses that are rarely observed in the logged data. Directly maximizing them can thus lead to harmful budget choices. ⇒ *We therefore need a selection rule that accounts for uncertainty when comparing candidate budgets and guards against harmful departures from the reference policy.*

In this paper, **we address both challenges by developing a sample-only framework for evaluating BoN policies and selecting the inference-time sampling budget N.** We proceed in two steps:

① **Estimation** (→ Section 4): To develop an estimator, we exploit a statistical property of BoN: the underlying order-statistic structure makes the otherwise unavailable density ratio between the BoN policy and the reference model expressible through the probabilities that a fresh reference-model sample scores below or at a given response, which can be estimated from samples alone. Building on this, we construct a doubly robust estimator of the BoN policy value (called BoN-DR) *without* access to response likelihoods. Here, BoN-DR separates the score used to *select* responses from the reward estimator used to *evaluate* them. Further, it reuses a shared auxiliary sample pool across candidate budgets and is unbiased for any finite pool size. We show that our BoN-DR yields asymptotically valid inference even when the reward estimator is misspecified, and that it attains the efficiency of an exact-ratio DR estimator as the pool grows and the reward estimator is consistent.

② **Selecting** N (→ Section 5): To select the sampling budget N, we develop two selection rules: (i) *Value-max* chooses the budget with the largest estimated policy value; and (ii) *Improvement-LCB* chooses the budget with the largest lower confidence bound (LCB) on the improvement over the reference policy, with a fallback to $N = 1$ if no LCB is positive. We show that both rules asymptotically attain the best policy value on the candidate grid, while Improvement-LCB additionally provides an asymptotic no-harm guarantee relative to the reference policy with high probability. To compute the LCB, we further construct simultaneous confidence bounds across candidate budgets. Unlike pointwise intervals, these bounds account for searching over the candidate grid and remain asymptotically valid after N is selected from the same data.

Our **contributions** are fourfold:[1] **(1)** We develop BoN-DR, a novel, unbiased doubly robust estimator for evaluating BoN policies from logged data, which requires only sample access to the reference model. **(2)** We prove statistical guarantees for BoN-DR, including robustness to reward estimator misspecification, asymptotic normality, and efficiency. **(3)** We propose two principled selection rules for N (i.e., Value-max and Improvement-LCB) with consistency and a no-harm guarantee. **(4)** We

[1]Code available at: `https://anonymous.4open.science/r/efficient_bon-C85B`.

evaluate our methods in synthetic experiments and on GSM8K across multiple reference and score models.

2 Related work[2].

BoN for inference-time alignment. BoN is widely used for inference-time alignment with reward or preference models (Stiennon et al., 2020; Nakano et al., 2021). Prior work studies the induced policy, its divergence to the reference policy, and the theoretical and empirical scaling behavior (e.g., Gui et al., 2024; Yang et al., 2024; Mroueh & Nitsure, 2025; Beirami et al., 2025; Sriraman & Block, 2026). Larger budgets can amplify proxy errors and cause reward overoptimization, motivating methods that modify the score, selection rule, or underlying model (Gao et al., 2023; Coste et al., 2024; Huang et al., 2025; Jinnai et al., 2025). We instead keep standard BoN fixed and study how to evaluate its downstream value and select its budget from logged data. Khalaf et al. (2025) consider a closely related problem, where they tune the BoN budget using proxy and true rewards of sampled responses. However, their method requires true rewards for all sampled candidates and does **not** provide uncertainty quantification. In contrast, we select N from logged data with only one observed reward per prompt and provide statistical guarantees.

Off-policy evaluation (OPE). Classical OPE estimates a target policy's value from data collected under a behavior policy using replay, importance weighting, or doubly robust estimation. (Li et al., 2011; Horvitz & Thompson, 1952; Strehl et al., 2010; Dudik et al., 2011). High-confidence OPE and safe policy improvement further use lower confidence bounds to guard against underperforming a baseline (Thomas et al., 2015a;b; Laroche et al., 2019). These methods cannot be directly applied to BoN under sample-only access because the required target-to-behavior density ratio depends on *unavailable* response likelihoods. We later exploit the order-statistic structure in BoN to express this ratio through score-rank probabilities, thereby enabling policy value estimation and uncertainty-aware budget selection from sampled responses and logged rewards.

3 Problem Setup

We consider a context $X \in \mathcal{X}$, an action $A \in \mathcal{A}$, and a scalar reward $Y \in \mathcal{Y}$. We call any (possibly stochastic) decision rule π mapping a context to an action a *policy*. We assume that the observed action A is generated by a *behavior* or *reference policy* $\pi_b(a \mid x) = \mathbb{P}(A = a|X = x)$. For example, in the LLM setting, π_b is induced by a *reference model* that generates a text A given a prompt X, based on which we obtain downstream signal Y (e.g., task success, human preference, click-through rate, etc.). We denote the reward function as $\mu(x, a) = \mathbb{E}[Y|X = x, A = a]$, and we observe an i.i.d. logged dataset $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$ from the population $Z = (X, A, Y) \sim \mathbb{P}$.

Inference-time alignment. We assume that we can *sample* from $\pi_b(\cdot \mid x)$ for any context x, but can neither access the model weights nor evaluate the response likelihood $\pi_b(a \mid x)$. This is typical for API-based models, which return sampled responses but no token-level probabilities. Our goal is to improve the deployed policy using only such samples and the logged data $\mathcal{D}$.

BoN policies. A well-established method for inference-time alignment is *best-of-N sampling* (BoN) (Nakano et al., 2021; Beirami et al., 2025). Here, a score function $s(x, a)$ first ranks actions a for each context x. For example, $s(x, a)$ could be an LLM judge (Zheng et al., 2023) or an estimator of the reward function $\mu(x, a)$. Then, for a fixed context $x_i \in \mathcal{X}$, BoN samples N i.i.d. actions $a_{i1}, \ldots, a_{iN} \sim \pi_b(\cdot \mid x_i)$ from the API and chooses the one with the largest score $a_N^{\star,s}(x_i) \in \arg\max_{j \le N} s(x_i, a_{i,j})$. The induced BoN policy is defined as $\pi_N^s(a \mid x_i) := \mathbb{P}\big(a_N^{\star,s}(x_i) = a \mid X = x_i\big)$. In particular, $\pi_1^s = \pi_b$, whereas increasing N progressively concentrates the policy on actions receiving high scores under s.

Policy evaluation. Our paper primarily focuses on the *evaluation* of the BoN policy π_N^s. For example, we might be interested in the performance of π_N^s for a specific N, or whether π_N^s for $N > 1$ improves over the base model π_b. A principled way to evaluate a policy π is via its *policy value* $V(\pi) = \mathbb{E}\Big[\sum_{a \in \mathcal{A}} \pi(a \mid X)\mu(X, a)\Big]$, which quantifies the expected reward under policy π (Kallus, 2018).

[2]We provide an extended related work in Appendix A

The choice of N. Evaluating $V(\pi_N^s)$ is directly tied to the choice of N, which controls how far π_N^s departs from π_b: a small N yields only marginal gains, while a large N concentrates the policy on high-scoring actions that are rare under π_b and thus weakly represented in $\mathcal{D}$. As a result, errors in the score function can be amplified, which can lead to reward overoptimization (Gao et al., 2023).

In this paper, we address two core challenges:

① **Off-policy evaluation** (→ Section 4): How can we efficiently estimate the value $V(\pi_N^s)$ of a BoN policy from data $\mathcal{D}$ generated by π_b when the reference model can only be sampled?
② **Selecting** N (→ Section 5): How can we use these estimates to select the sampling budget N with statistical guarantees?

4 Efficient Off-Policy Evaluation for BoN

In this section, we develop an estimator for the value of the BoN policy $V(\pi_N^s)$. A natural starting point is the doubly robust (DR) estimator, which is widely used for off-policy evaluation due to its favorable statistical properties (Robins et al., 1994; Dudik et al., 2011). For a fixed BoN policy π_N^s, the DR estimator is

$$\hat{V}_{\mathrm{DR}}(\pi_N^s) = \frac{1}{n}\sum_{i=1}^{n} \psi_N(z_i), \psi_N(z_i) = \sum_{a\in\mathcal{A}} \pi_N^s(a \mid x_i)\hat{\mu}(x_i,a) + \frac{\pi_N^s(a_i \mid x_i)}{\pi_b(a_i \mid x_i)}\big(y_i - \hat{\mu}(x_i,a_i)\big), \quad (1)$$

where $z_i = (x_i, a_i, y_i)$ and $\hat{\mu}$ is an estimator of the reward function μ. The plug-in term (first term in the sum) evaluates the target policy under the estimated reward function, while the correction term (second term) uses the observed outcome to correct for errors in the estimated reward function, weighted by the density ratio between the BoN and behavior policies.

Why the standard DR estimator is not applicable. Direct application of the DR estimator in the form of (1) is *not* possible in the inference-time alignment setting due to two main challenges: (i) both terms require knowledge of $\pi_N^s(a \mid x_i)$, but the induced BoN probabilities are not available from a black-box sampler. Attempting to estimate $\pi_N^s(a \mid x_i)$ from the data is extremely challenging in the enormous and sparse text action space $\mathcal{A}$. (ii) The correction term requires the density ratio $w_N(x,a) := \frac{\pi_N^s(a|x)}{\pi_b(a|x)}$. Because the action space $\mathcal{A}$ consists of high-dimensional text sequences, individual sequence probabilities can be extremely small, and density-ratio estimation can suffer from high variance and numerical instability.

To address this, we first establish an alternative representation of the DR score for the inference-time alignment setting (Section 4.1). We then construct a pooled estimator that reuses auxiliary samples across candidate budgets (Section 4.2) and establish theoretical guarantees (Section 4.3).

4.1 Inference-time DR score

Our key observation is that the target policy is not arbitrary: π_N^s is obtained by applying a known Best-of-N selection rule to samples from π_b. We can therefore exploit two pieces of structure jointly: (i) although π_b cannot be evaluated, we can sample from $\pi_b(\cdot \mid x)$ at inference time, and (ii) the construction of the BoN policy imposes a particular relationship between π_N^s and π_b.

In the following, we show that DR score admits a specific representation only in terms of one-dimensional score ranks under the behavior policy. For fixed x and action a, we define the strict and weak cumulative density functions (CDFs) of the score via

$$F_x^-(a) = \mathbb{P}_{A\sim\pi_b(\cdot|x)}\left(s(x,A) < s(x,a)\right) \text{ and } F_x(a) = \mathbb{P}_{A\sim\pi_b(\cdot|x)}\left(s(x,A) \le s(x,a)\right). \quad (2)$$

Thus, $F_x^-(a)$ and $F_x(a)$ describe the rank of a according to the score s among actions sampled from the behavior policy. Importantly, both quantities are probabilities of simple comparison events and can therefore be estimated using samples from $\pi_b(\cdot \mid x)$ without evaluating any sequence likelihoods. The following proposition shows that these rank probabilities are sufficient to express the entire DR score for a BoN policy (which we call BoN-DR score).

Proposition 4.1 (BoN-DR score)**.** *Suppose that distinct actions have distinct scores, i.e., $s(x,a) = s(x,a') \implies a = a'$, for any $a, a' \in \mathcal{A}$. Define $g_N(u,v) = \sum_{\ell=0}^{N-1} u^{N-1-\ell}v^\ell$. Then, for $z = (x,a,y)$ with $\pi_b(a \mid x) > 0$, the DR score for the BoN policy can be written as*

$$\psi_N(z) = \mathbb{E}_{A'\sim\pi_b(\cdot|x)}\left[g_N\big(F_x(A'), F_x^-(A')\big)\,\hat{\mu}(x,A')\right] + g_N\big(F_x(a), F_x^-(a)\big)\,\big(y - \hat{\mu}(x,a)\big). \quad (3)$$

Proof. This follows by combining the BoN order-statistic identity in Beirami et al. (2025) with the DR score formula in Eq. (1). We refer to Appendix B for details. □

Proposition 4.1 resolves the two computational obstacles that we identified above: (i) it eliminates the need to evaluate either $\pi_b(a \mid x)$ or $\pi_N^s(a \mid x)$; and (ii) the density ratio is replaced by $g_N(F_x(a), F_x^-(a))$, which depends only on the relative rank of a under the score distribution induced by the behavior policy. Consequently, apart from evaluating the score function s and outcome model $\hat{\mu}$, every quantity in Eq. (3) can be approximated using only samples from the behavior policy. In particular, $F_x(a)$ and $F_x^-(a)$ are expectations of bounded indicator variables, e.g., $F_x(a) = \mathbb{E}_{A' \sim \pi_b(\cdot|x)}[\mathbf{1}\{s(x, A') \leq s(x, a)\}]$, thus turning the original high-dimensional density-ratio evaluation problem into the estimation of simple rank probabilities.

The remaining challenge is therefore how we should estimate the expectations and score-rank functionals in Eq. (3) given a finite number of auxiliary samples from $\pi_b(\cdot \mid x)$. In the next section, we show that their particular polynomial structure admits natural unbiased U-statistic estimators.

4.2 Pooled BoN-DR estimator

We now construct an estimator of the DR score in Eq. (1) using a finite, reusable pool of samples from the behavior policy. Proposition 4.1 shows that the DR score can be estimated using only samples from the behavior policy. A direct Monte Carlo implementation of Eq. (3) could draw fresh samples from the behavior policy each time an expectation or score rank needs to be estimated, and repeat this procedure separately for every candidate value of N. For generative models, however, sampling may be computationally costly.

Instead, we propose to draw a single reusable pool of samples for each context and use it to estimate both components of the DR score (i.e., the plug-in term and the correction term), as well as the values of other BoN policies with different N. For each context x_i, we draw one pool of behavior-policy samples $\tilde{a}_i = (\tilde{a}_{i1}, \ldots, \tilde{a}_{iM}) \overset{\text{iid}}{\sim} \pi_b(\cdot \mid x_i)$, where M is a user-chosen computational budget, and let us denote $\bar{a}_i = (a_i, \tilde{a}_{i1}, \ldots, \tilde{a}_{iM}) \in \mathcal{A}^{M+1}$ the vector of Monte Carlo samples and behavior actions.

To derive our BoN-DR estimator, we leverage *U-statistics* (van der Vaart, 1998), which estimate a quantity defined as an expectation over r i.i.d. draws via a single larger pool of M draws. The basic idea is to estimate the quantity by averaging the same quantity over all size-r subsets of the pool. Each subset has the correct distribution, so the resulting estimator is unbiased while reusing all available samples. Below, we apply this construction separately to (i) the plug-in term and (ii) correction term of the DR score and then combine the two into our pooled BoN-DR estimator.

Estimator for the plug-in term. By Proposition 4.1, the plug-in term can be written as

$$v_N^s(x_i) = \mathbb{E}_{A \sim \pi_N^s(\cdot|x_i)}[\hat{\mu}(x_i, A)] = \mathbb{E}\left[\hat{\mu}(x_i, A_N^{\star,s}) \mid X_i = x_i\right], \tag{4}$$

where $A_N^{\star,s}$ is the action selected from N i.i.d. draws from $\pi_b(\cdot \mid x_i)$ according to the score s. Since $\bar{a}_i$ contains $M+1$ such draws, we estimate $v_{i,N}^s$ by averaging the fitted reward of the BoN winner over all size-N subsets:

$$\hat{v}_{i,N}^{s,M+1} = \binom{M+1}{N}^{-1} \sum_{J \subseteq \{0,\ldots,M\}, |J|=N} \hat{\mu}\left(x_i, \bar{a}_{i,J}^{\star,s}\right). \tag{5}$$

This estimator can be evaluated without enumerating all subsets. Let $\mathcal{G}_i$ index the unique actions appearing in $\bar{a}_i$, and let $\bar{a}_{ig}$ denote the corresponding action. We define

$$L_{ig} = \sum_{m=0}^{M} \mathbf{1}\{s(x_i, \bar{a}_{im}) < s(x_i, \bar{a}_{ig})\} \text{ and } E_{ig} = \sum_{m=0}^{M} \mathbf{1}\{s(x_i, \bar{a}_{im}) \leq s(x_i, \bar{a}_{ig})\}. \tag{6}$$

Then, an alternative estimator is

$$\hat{v}_{i,N}^{s,M+1}(x_i, \bar{a}_i) = \sum_{g \in \mathcal{G}_i} \hat{\mu}(x_i, \bar{a}_{ig}) \frac{\binom{E_{ig}}{N} - \binom{L_{ig}}{N}}{\binom{M+1}{N}}. \tag{7}$$

Indeed, the numerator counts exactly the size-N subsets for which $\bar{a}_{ig}$ is the highest-scoring action. We use the convention $\binom{u}{N} = 0$ for $u < N$.

Estimator for the correction term. For the correction term, we treat the logged action a_i as fixed and use the M auxiliary samples to estimate $w_N^s(x_i, a_i) = g_N(F_{x_i}(a_i), F_{x_i}^-(a_i))$. We define

$$L_i = \sum_{m=1}^{M} \mathbf{1}\left\{s(x_i, \tilde{a}_{im}) < s(x_i, a_i)\right\}, \text{ and } E_i = \sum_{m=1}^{M} \mathbf{1}\left\{s(x_i, \tilde{a}_{im}) \le s(x_i, a_i)\right\}. \quad (8)$$

Under a no-score-ties assumption, $E_i - L_i$ counts the number of auxiliary draws equal to a_i. By averaging over all size-$(N-1)$ subsets of the auxiliary pool, we yield the estimator

$$\hat{w}_{i,N}^{s,M}(x_i, \bar{a}_i) = N \sum_{t=0}^{\min\{E_i - L_i, N-1\}} \frac{1}{t+1} \frac{\binom{E_i - L_i}{t}\binom{L_i}{N-1-t}}{\binom{M}{N-1}}. \quad (9)$$

Here, t denotes the number of additional copies of a_i among the $N-1$ sampled competitors, while the factor $1/(t+1)$ accounts for the $t+1$ identical occurrences of the winning action.

Pooled BoN-DR estimator. We define our **pooled BoN-DR estimator** as

$$\psi_N^M(z_i, \tilde{a}_i) = \hat{v}_{i,N}^{s,M+1}(x_i, \bar{a}_i) + \hat{w}_{i,N}^{s,M}(x_i, \bar{a}_i)\,(y_i - \hat{\mu}(x_i, a_i)), \quad (10)$$

$$\hat{V}_{\mathrm{DR}}^{M}(\pi_N^s) = \frac{1}{n}\sum_{i=1}^{n} \psi_N^M(z_i, \tilde{a}_i). \quad (11)$$

Thus, a single pool of M additional generations per context is sufficient to estimate both components of the DR score. Moreover, the same pool can be reused across different values of N, which is particularly important for the policy-selection problem considered in the following section.

4.3 Theoretical guarantees

We now establish asymptotic guarantees for our pooled BoN-DR estimator, including valid inference and efficiency. For this, we assume that s and $\hat{\mu}$ are constructed independently of $\mathcal{D}$ (e.g., external models or a sample split), together with standard sampling and moment conditions (Assumption B.1).

Theorem 4.2 (Validity and efficiency of the pooled BoN-DR estimator)**.** *Fix N and condition on the training sample used to construct s and $\hat{\mu}$.*

***(a)** Finite-M validity. For any fixed $M \ge N-1$, $\sqrt{n}\left(\hat{V}_{\mathrm{DR}}^{M}(\pi_N^s) - V(\pi_N^s)\right) \rightsquigarrow \mathcal{N}(0, \sigma_{N,M}^2)$, where $\sigma_{N,M}^2 = \mathrm{Var}\left(\psi_N^M(Z, \tilde{A})\right)$. This result does not require $\hat{\mu}$ to be a consistent estimator of μ.*

***(b)** Infinite-M efficiency. Suppose additionally that $M = M_n \to \infty$ and that the reward function estimator $\hat{\mu}$ is consistent for the true reward function μ in L_2. Then $\sqrt{n}\left(\hat{V}_{\mathrm{DR}}^{M_n}(\pi_N^s) - V(\pi_N^s)\right) \rightsquigarrow \mathcal{N}(0, \sigma_{N,\mathrm{eff}}^2)$, where $\sigma_{N,\mathrm{eff}}^2$ is the variance of the efficient influence function $\phi_N^{\mathrm{eff}}(Z) = \sum_{a\in\mathcal{A}} \pi_N^s(a \mid X)\mu(X,a) + w_N^s(X,A)\big(Y - \mu(X,A)\big) - V(\pi_N^s)$.*

Thus, conditional on the learned policy and behavior sampler, the proposed estimator attains the same first-order efficiency bound as an oracle DR estimator with access to the exact BoN density ratio. See Appendix B for the proof.

Part **(a)** is particularly important for inference-time alignment. Valid inference does *not* require the reward estimator to approximate the true conditional mean. For example, s and $\hat{\mu}$ may be given by an LLM judge that is useful for ranking responses but is *not* a statistically consistent model of $\mathbb{E}[Y \mid X, A]$. The DR correction still yields a consistent, root-n estimator of the value of the resulting BoN policy. Part **(b)** adds that, if $\hat{\mu}$ is additionally consistent, the BoN-DR estimator attains the corresponding first-order efficiency bound.

We next clarify how the auxiliary sampling budget M affects the estimator. Because the pooled estimators are unbiased for every finite M, increasing M removes Monte Carlo variance rather than a first-order bias. For fixed N, under the boundedness conditions above, it is sufficient that $M_n \to \infty$; no particular rate of growth relative to n is required.

5 Selecting the optimal sampling budget N for BoN

Our goal is now to select the optimal sampling budget N_j from a candidate grid $\mathcal{N} = \{N_1, \ldots, N_J\}$. In the following, we introduce two selection rules: **(i) Value-max**, which selects the budget with the

largest estimated policy value, and **(ii) Improvement-LCB**, which selects the budget with the largest lower confidence bound (LCB) on improvement over the reference policy and falls back to $N = 1$ when no positive improvement is supported by the data. The latter thus guarantees improvement over the base model with high probability.

5.1 VALUE-MAX

Value-max selects the budget with the largest estimated policy value:

$$\hat{N}_{\text{value}} \in \arg\max_{N \in \mathcal{N}} \hat{V}_{\text{DR}}^{M}(\pi_N^s). \tag{12}$$

The following corollary shows that, over a fixed candidate grid, this rule asymptotically recovers the best policy value in the class.

Corollary 5.1 (Oracle consistency of value maximization)**.** *Suppose that the conditions of Theorem 4.2(a) hold uniformly over the finite candidate grid $\mathcal{N}$. Then, conditional on the training sample, $V(\pi^s_{\hat{N}_{\text{value}}}) \xrightarrow{p} \max_{N\in\mathcal{N}} V(\pi_N^s)$. The proof is in Appendix B.*

Importantly, this conclusion inherits the robustness of the BoN-DR estimator: it does *not* require $\hat{\mu}$ to consistently estimate the true reward function. However, the result is an *oracle optimality* guarantee, and, in finite samples, estimation noise can still cause $\hat{N}_{\text{value}}$ to select a policy for which the value is below $V(\pi_b)$. This motivates our selection rule for safe improvement next.

5.2 IMPROVEMENT-LCB

To obtain a no-harm guarantee relative to the behavior policy, we now formulate the selection rule directly in terms of improvement over the behavior policy. To do so, we define

$$\Delta_N := V(\pi_N^s) - V(\pi_1^s) = V(\pi_N^s) - V(\pi_b) \tag{13}$$

for each $N \in \mathcal{N} \setminus \{1\}$. *Improvement-LCB* then selects a candidate $N > 1$ only when its simultaneous lower confidence bound for Δ_N is positive; otherwise, we fall back to $N = 1$. We use simultaneous rather than pointwise bounds because N is selected adaptively from the same candidate grid.

• To construct the LCB, we estimate each candidate policy's improvement over the reference policy and its sampling variability. To do so, for each $N \in \mathcal{N}\setminus\{1\}$, we define the doubly robust improvement score $D_{i,N}^M := \psi_N^M(z_i, \tilde{a}_i) - \psi_1^M(z_i, \tilde{a}_i)$, and the corresponding estimator $\hat{\Delta}_N^M := \frac{1}{n}\sum_{i=1}^n D_{i,N}^M = \hat{V}_{\text{DR}}^M(\pi_N^s) - \hat{V}_{\text{DR}}^M(\pi_1^s)$. The corresponding variance is estimated by $\hat{\tau}_{N,M}^2 := \frac{1}{n}\sum_{i=1}^n \left(D_{i,N}^M - \hat{\Delta}_N^M\right)^2$.

• To obtain *simultaneous* LCBs across candidate budgets, we need to account for the dependence induced by reusing the same evaluation data and auxiliary samples pools across all $N \in \mathcal{N}$. We do so by using a joint multiplier bootstrap (Chernozhukov et al., 2013). Conditional on the sample, we thus draw independent multipliers ξ_i satisfying $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}[\xi_i^2] = 1$, such as standard Gaussian variables. Formally, we define

$$T_\Delta^{*,+} := \max_{N\in\mathcal{N}\setminus\{1\}} \frac{n^{-1/2}\sum_{i=1}^n \xi_i \left(D_{i,N}^M - \hat{\Delta}_N^M\right)}{\hat{\tau}_{N,M}}. \tag{14}$$

Let $\hat{c}_{1-\alpha}^{\Delta,+}$ denote the conditional $(1-\alpha)$-quantile of $T_\Delta^{*,+}$. Thus, simultaneous LCB is given by

$$\widehat{\text{LCB}}_\Delta(N) := \hat{\Delta}_N^M - \hat{c}_{1-\alpha}^{\Delta,+}\frac{\hat{\tau}_{N,M}}{\sqrt{n}}, \qquad N \in \mathcal{N}\setminus\{1\}. \tag{15}$$

Improvement-LCB selects the candidate with the strongest LCB on improvement over the reference policy and falls back to $N = 1$ when no candidate has a positive LCB:

$$\hat{N}_{\text{LCB}} := \begin{cases} \arg\max_{N\in\mathcal{N}\setminus\{1\}} \widehat{\text{LCB}}_\Delta(N), & \max_{N\in\mathcal{N}\setminus\{1\}} \widehat{\text{LCB}}_\Delta(N) > 0, \\ 1, & \text{otherwise.} \end{cases} \tag{16}$$

Theorem 5.2 (Safe improvement ("no-harm guarantee"))**.** *Suppose that the conditions of Theorem 4.2(a) hold uniformly over the finite candidate grid $\mathcal{N}$ and that the variances $\tau_{N,M}^2$ are uniformly bounded away from zero. Conditional on the training sample, it follows that $V(\pi^s_{\hat{N}_{\text{LCB}}}) \xrightarrow{p} \max_{N\in\mathcal{N}} V(\pi_N^s)$ and $\mathbb{P}\left(V(\pi^s_{\hat{N}_{\text{LCB}}}) \geq V(\pi_b)\right) \geq 1 - \alpha - o(1)$. Proof in Appendix B.*

Theorem 5.2 formalizes the no-harm property of the Improvement-LCB selection rule. A candidate $N > 1$ is selected *only* when the improvement over the behavior policy is supported by the simultaneous LCB; otherwise, the rule defaults to $N = 1$. The guarantee combines asymptotic no-harm with consistency; i.e., as the sample size grows, the conservatism induced by the LCB vanishes, and Improvement-LCB recovers the best policy value over the fixed candidate grid. In finite samples, Improvement-LCB may select smaller budgets than Value-max because it explicitly accounts for estimation uncertainty.

6 Experiments

The purpose of our experiments is primarily to verify our theoretical properties. RQ1 We use synthetic experiments to test whether our BoN-DR estimator remains **robust to reward estimator misspecification** while benefiting from informative reward predictions. RQ2 We assess whether the estimators accurately **recover BoN policy values** across different reference models and score function regimes in realistic LLM settings. RQ3 We test whether Improvement-LCB reduces harmful choices of N under reward overoptimization and thus leads to **safe inference-time budget selection**.

Settings. We consider two settings: **(1) Synthetic experiments** with known ground truth, following standard practice in the OPE literature (Dudik et al., 2011; Saito & Joachims, 2022). Here, the score function errors and reward estimator errors can be varied independently through parameters γ (which increases the scores of rarely sampled, low-quality actions and thereby changes the induced BoN policy) and δ (which introduces errors in the predicted rewards without changing the induced policy). Appendix C.1 provides details. **(2) Realistic LLM evaluation setting** using GSM8K (Cobbe et al., 2021), with three reference models (i.e., Gemma-2-2B-IT (Team et al., 2024), Llama-3.2-3B-Instruct (Grattafiori et al., 2024), and Qwen3.5-4B) and two external reward models as score functions (i.e., OASST-RM-2.1-Pythia-1.4B (Köpf et al., 2023) and Skywork-Reward-V2-Qwen3-1.7B). Details are in Appendix C.2. The true reward is the exact-match correctness (see Appendix C.5). Each prompt has a stored pool of 512 responses. In each evaluation, one response is treated as the logged observation where the reward is revealed, while the remaining responses form the auxiliary pool (see Appendix D.3).

Baselines.[3] To our knowledge, there are **no** existing baselines that directly address OPE estimation for BoN policies from logged data. Instead, we compare three estimators based on Eq. (11): **(i) Plug-in** averages predictions of reward estimator under the target policy; **(ii) BoN-IPW** weights logged rewards using estimated rank probabilities; and **(iii) BoN-DR** uses an additional doubly robust correction. All estimators use the *same* logged data and auxiliary samples. For selection, we use the rules from Section 5; i.e., **Value-max** and **Improvement-LCB**.

Performance metrics. We measure bias, RMSE, interval width, and coverage of nominal 95% pointwise confidence intervals. Selection is assessed by mean gain over $N = 1$ and harm rate, the fraction of repetitions with negative gain.



Figure 2: **Robust to reward estimator misspecification.** Estimated BoN policy value at $N = 256$. We fix $\gamma = 0.65$ and vary δ from 0 to 0.6 to increase the reward estimator error. Shown: **(a)** Absolute bias of policy value estimation. **(b)** Empirical coverage of the nominal 95% pointwise CI, the dotted line shows the nominal level. **(c)** Gain of the budget selected by Improvement-LCB, where the dotted line indicates no improvement over $N = 1$.

N	Estimator	|Bias|	CI width	Coverage
8	Plug-in	0.049	0.003	0.000
	BoN-IPW	0.002	0.097	0.940
	BoN-DR	**0.001**	0.052	**0.947**
256	Plug-in	0.369	0.004	0.000
	BoN-IPW	**0.002**	0.377	0.887
	BoN-DR	0.003	0.455	**0.940**

* CI width and empirical coverage are computed for nominal 95% pointwise CIs.

Table 1: **Sensitivity to sampling budget N.** Here: $\gamma = 0.65$, $\delta = 0.4$. **Bold** shows the lowest absolute bias and the coverage closest to 95% for each N.

RQ1 **Robustness to reward estimator misspecification.** We evaluate robustness to reward estimator misspecification by varying δ, which introduces errors in the reward estimator without changing the

[3] Additional details on the BoN-IPW and plug-in estimators are provided in Appendix C.3.

induced policy. We compare Plug-in, BoN-IPW, and BoN-DR in terms of absolute bias, CI coverage, and policy value gain over $N = 1$. Results are in Figure 2. Plug-in becomes biased as the errors in the reward prediction are increased. In contrast, BoN-DR and BoN-IPW remain nearly unbiased as δ increases. Further, they maintain CI coverage close to the nominal 95% level across the considered levels of misspecification. ⇒ *As suggested by our theory, BoN-DR remains robust to reward model misspecification while supporting valid CIs.*

We next examine whether these results persist across different BoN budgets by comparing $N = 8$ and $N = 256$ in Table 1. ⇒ *The results again show that the benefits of BoN-DR are robust.*

In Appendix D.2–D.3, we provide additional sensitivity analyses on the size of the validation set and the Monte Carlo pool. We study how the amount of validation data affects downstream policy value estimation and budget selection, and show that increasing the Monte Carlo pool reduces the additional Monte Carlo variance of BoN-DR and brings it closer to the efficiency bound under a correctly specified reward estimator.

RQ2 Performance in recovering BoN policy values in realistic LLM settings. Figure 3 compares the estimated policy values from Plug-in, BoN-IPW, and BoN-DR with the oracle policy values under different configurations. BoN-DR shows the most accurate estimates. Moreover, at $N = 256$, BoN-DR yields substantially narrower CIs, with BoN-IPW intervals being 1.7–5.5× wider. Notably, the Plug-in estimator remains biased even though the score function ranks responses accurately in the Skywork setting. This illustrates that effective ranking does not necessarily translate into accurate policy value estimation. ⇒ *BoN-DR accurately recovers BoN policy values in real-world settings while providing substantially more precise estimates than BoN-IPW.*

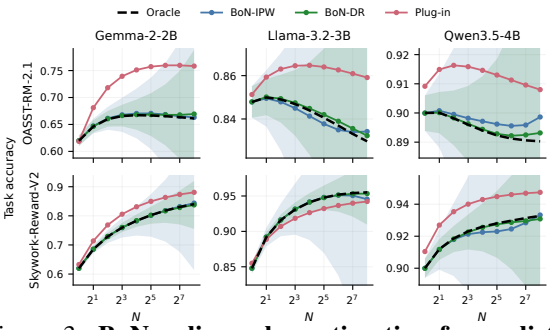


Figure 3: **BoN policy value estimation for realistic LLM setting on GSM8K.** Rows are external reward models used as score functions, and columns are reference models. **Black dashed:** oracle benchmark policy value. **Lines:** mean estimated policy value across repetitions. **Shaded:** 80% pointwise CI. Plug-in CIs are not faithful wrt. nominal coverage and thus omitted.

We provide additional results for CI coverage, error decomposition, overlap at large budgets, and robustness to alternative reward predictors in Appendices D.4–D.7.

RQ3 Safe inference-time budget selection ("no-harm guarantee"). We study whether Improvement-LCB, which explicitly accounts for estimation uncertainty, helps avoid harmful choices of N, i.e., budgets that reduce policy value relative to the reference policy. This is particularly important under reward overoptimization with a strong reference model, where larger budgets achieve higher scores but can lower downstream performance. For comparison, we also report a Naïve approach here (which chooses the response with the highest score at a fixed budget).

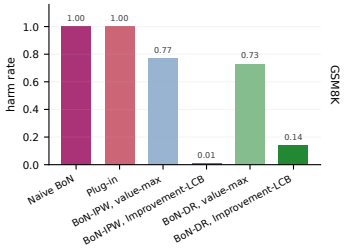


Figure 4: **Comparison of budget selection rules for Qwen3.5-4B with OASST-RM-2.1.** Here, the harm rate is defined as the fraction of evaluations with negative gain relative to $N = 1$. Lower is better. Plug-in uses value-max, and Naïve BoN selects $N_{\max} = 256$.

We have two main findings (see Figure 4). First, both Naïve BoN and Plug-in perform poorly overall and often lead to harmful selections. In contrast, our proposed selection rules (=Value-max and Improvement-LCB) are substantially better. Second, Improvement-LCB leads to fewer harmful selections than Value-max. Appendix E reports detailed results across various configurations. ⇒ *Thanks to the uncertainty-aware approach, Improvement-LCB can reduce harmful selections compared to Value-max, which is in line with our no-harm guarantee.*

Conclusion. We present, to our knowledge, the first framework for evaluating and selecting BoN policies under sample-only access, with valid inference and a no-harm guarantee. As a **limitation**, it requires logged data and loses precision at large budgets. Future work could use our per-prompt scores to learn prompt-specific budgets and extend likelihood-free evaluation to other inference-time methods built on black-box models.

AI USE STATEMENT

The research questions and core methodological ideas were developed by the authors. Generative AI tools were used as research assistants for literature retrieval, feedback on research and experimental design, implementation and debugging, synthetic-data and figure preparation, checking and refining mathematical claims and proofs, interpreting results, and drafting and editing parts of the manuscript. All AI-assisted outputs were reviewed and, where applicable, verified against the implementation, experiments, literature, or mathematical derivations; the authors made all final decisions and take responsibility for the content. LLM use as part of the experimental data-generation pipelines is described separately in the corresponding experimental appendices.

REPRODUCIBILITY STATEMENT

Assumptions and proofs are provided in Appendix B. The synthetic data-generating process, the GSM8K setup, the estimators, and the evaluation protocols are described in Appendix C. The models and datasets are publicly available, and the code is available at `https://anonymous.4open.science/r/efficient_bon-C85B`.

ETHICS STATEMENT

Our experiments use a public benchmark, publicly available open-weight models, and synthetic data; no human subjects were involved and no policies were deployed to users. Our methods aim to make inference-time alignment safer by requiring statistical evidence of improvement over the reference model before a larger sampling budget is deployed. However, the no-harm guarantee is asymptotic and concerns the average value under the chosen reward; it does not rule out harm on individual prompts or subgroups, nor along dimensions that the reward does not capture. Applications that evaluate policies on logged user feedback should additionally consider privacy, and the choice of reward should account for factual accuracy, safety, and fairness alongside the optimized outcome.

REFERENCES

Gholamali Aminian, Idan Shenfeld, Amir Reza Asadi, Ahmad Beirami, and Youssef Mroueh. Best-of-n through the smoothing lens: Kl divergence and regret analysis. In *International Conference on Learning Representations*, volume 2026, pp. 105057–105090, 2026.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3580–3602. PMLR, 13–19 Jul 2025.

Aniruddha Bhargava, Lalit Jain, Branislav Kveton, Ge Liu, and Subhojyoti Mukherjee. Off-policy evaluation from logged human feedback. *arXiv preprint arXiv:2406.10030*, 2024.

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

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, pp. 2786–2819, 2013.

Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Aviral Kumar, Rishabh Agarwal, Sridhar Thiagarajan, Craig Boutilier, and Aleksandra Faust. Inference-aware fine-tuning for best-of-n sampling in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 78936–78959, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, volume 2024, pp. 50905–50931, 2024.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *ICML*, 2011.

Dennis Frauen, Athiya Deviyani, Mihaela van der Schaar, and Stefan Feuerriegel. Nonparametric llm evaluation from preference data. In *International Conference on Machine Learning*, 2026.

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. *arXiv preprint arXiv:2407.21783*, 2024.

Lin Gui, Cristina Gârbacea, and Victor Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *Advances in Neural Information Processing Systems*, 37: 2851–2885, 2024.

Tobias Hatt, Daniel Tschernutter, and Stefan Feuerriegel. Generalizing off-policy learning under sample selection bias. In *UAI*, 2022.

Konstantin Hess, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Efficient and sharp off-policy learning under unobserved confounding. In *International Conference on Learning Representations*, volume 2026, pp. 57552–57587, 2026.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

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 *International Conference on Machine Learning*, pp. 25075–25126. PMLR, 2025.

Ying Jin, Zhimei Ren, Zhuoran Yang, and Zhaoran Wang. Policy learning "without" overlap: Pessimism and generalized empirical bernstein's inequality. *The Annals of Statistics*, 53(4): 1483–1512, 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.

Nathan Kallus. Balanced policy evaluation and learning. In *NeurIPS*, 2018.

Nathan Kallus. More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association*, 116(534):646–658, 2021. doi: 10.1080/01621459.2020.1788948.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21:1–63, 2020.

Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2021.

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.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

Manuel Kosmas, Anne-Sophie Mayer, Stefan Feuerriegel, and Frank Bodendorf. From code to culture: Aligning large language models with corporate values. *Available at SSRN 5615933*, 2025.

Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pp. 640–648. PMLR, 2021.

Romain Laroche, Paul Trichelair, and Remi Tachet Des Combes. Safe policy improvement with baseline bootstrapping. In *International conference on machine learning*, pp. 3652–3661. PMLR, 2019.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Zihao Li, Xiang Ji, Minshuo Chen, and Mengdi Wang. Policy evaluation for reinforcement learning from human feedback: A sample complexity analysis. In *International Conference on Artificial Intelligence and Statistics*, pp. 2737–2745. PMLR, 2024.

Youssef Mroueh and Apoorva Nitsure. Information theoretic guarantees for policy alignment in large language models. *Transactions on Machine Learning Research*, 2025.

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

Yusuke Narita, Shota Yasui, and Kohei Yata. Efficient counterfactual learning from bandit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33(01), pp. 4634–4641, 2019.

Vinod Raman, Hilal Asi, and Satyen Kale. Adabon: Adaptive best-of-n alignment. *arXiv preprint arXiv:2505.12050*, 2025.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

J Jon Ryu, Jeongyeol Kwon, Benjamin Koppe, and Kwang-Sung Jun. Improved offline contextual bandits with second-order bounds: Betting and freezing. *arXiv preprint arXiv:2502.10826*, 2025.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. In *International Conference on Machine Learning (ICML)*, 2022.

Jonas Schweisthal, Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Reliable off-policy learning for dosage combinations. In *NeurIPS*, 2023.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot-Desenonges, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shahriari, Sarah Perrin, Abram Friesen, Geoffrey Cideron, et al. Bond: Aligning llms with best-of-n distillation. In *International Conference on Learning Representations*, volume 2025, pp. 59040–59060, 2025.

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.

Ved Sriraman and Adam Block. Revisiting the (sub) optimality of best-of-n for inference-time alignment. *arXiv preprint arXiv:2603.05739*, 2026.

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.

Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. *Advances in neural information processing systems*, 23, 2010.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29(01), 2015a.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388. PMLR, 2015b.

Aart van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 1998. ISBN 0521496039.

Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio P Calmon. Soft best-of-n sampling for model alignment. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6. IEEE, 2025.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pp. 3589–3597. PMLR, 2017.

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

Zhuohao Yu, Zhiwei Steven Wu, and Adam Block. From curiosity to caution: Mitigating reward hacking for best-of-n with pessimism. *arXiv preprint arXiv:2604.04648*, 2026.

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

A Extended related work

BoN and inference-time compute. Theoretical work characterizes the BoN policy, its KL divergence to the reference policy, and its optimality and scaling behavior (Gui et al., 2024; Yang et al., 2024; Mroueh & Nitsure, 2025; Beirami et al., 2025; Huang et al., 2025; Sriraman & Block, 2026). More broadly, repeated sampling with a verifier or reward model is a central way to scale inference-time compute (Brown et al., 2024; Snell et al., 2024), where best-of-k performance is commonly estimated via the unbiased pass@k estimator (Chen et al., 2021). Our plug-in term (Eq. (5)) follows the same U-statistic principle, but such estimators require the reward of every sampled response. In our setting, rewards are observed only for logged responses, which is why BoN-DR needs its correction term.

Mitigating reward overoptimization. Existing mitigations change the deployed policy by modifying (i) the score, e.g., via reward-model ensembles (Coste et al., 2024) or uncertainty penalties (Yu et al., 2026); (ii) the selection rule, e.g., via regularized, pessimistic, or soft BoN (Jinnai et al., 2025; Huang et al., 2025; Verdun et al., 2025; Aminian et al., 2026); or (iii) the model, via fine-tuning or distillation (Gui et al., 2024; Sessa et al., 2025; Chow et al., 2025). Some of these additionally require response likelihoods or model weights. Our work is complementary: we keep standard BoN fixed and evaluate it and select its budget from logged data.

Selecting inference-time parameters. Closest to our work, Khalaf et al. (2025) tune the BoN budget via the peak of the true reward. Their method is sample-only but requires true rewards for all sampled responses and provides no uncertainty quantification. Raman et al. (2025) instead allocate a given budget adaptively across prompts. In contrast, we select N from logged data with simultaneous confidence bounds and a no-harm guarantee.

Off-policy evaluation. Beyond the classical estimators (Li et al., 2011; Horvitz & Thompson, 1952; Strehl et al., 2010; Dudik et al., 2011), OPE has been extended to improve efficiency (Wang et al., 2017; Narita et al., 2019; Kallus & Uehara, 2020), handle limited overlap or distribution shift (Kallus, 2018; 2021; Schweisthal et al., 2023; Hatt et al., 2022; Hess et al., 2026), and large action spaces via embeddings (Saito & Joachims, 2022), as well as to human-feedback and preference data (Li et al., 2024; Bhargava et al., 2024; Frauen et al., 2026). All of these require the action probabilities of the behavior policy, which are unavailable for free-form text under sample-only access. We instead exploit the known BoN selection mechanism, which expresses the density ratio through rank probabilities.

Uncertainty-aware policy selection. High-confidence OPE and safe policy improvement use lower confidence bounds to avoid underperforming a baseline (Thomas et al., 2015a;b; Laroche et al., 2019), with related confidence-based selection methods for contextual bandits (Kuzborskij et al., 2021; Karampatziakis et al., 2021; Ryu et al., 2025; Jin et al., 2025). Improvement-LCB applies this principle to BoN, with simultaneous bounds across budgets that share the same logged data and auxiliary samples.

B PROOFS

B.1 REGULARITY CONDITIONS

We first make precise the regularity conditions used in the asymptotic results. For the efficiency result, let s_n and $\hat{\mu}_n$ denote the fitted score and reward functions at sample size n; the main text suppresses this dependence on n. These functions may coincide.

Let $\mathcal{T}_n$ denote the data used to fit $(s_n, \hat{\mu}_n)$.

Assumption B.1 (Regularity conditions)**.** The following conditions hold.

(i) **Sampling.** The evaluation observations

$$Z_i = (X_i, A_i, Y_i), \qquad i = 1, \dots, n,$$

are i.i.d., with

$$A_i \mid X_i = x \sim \pi_b(\cdot \mid x), \qquad \mathbb{E}[Y_i \mid X_i = x, A_i = a] = \mu(x, a).$$

The action space $\mathcal{A}$ is finite or countable. The training data $\mathcal{T}_n$ are independent of the evaluation sample.

Conditional on X_i, the auxiliary actions

$$\tilde{A}_{i1}, \dots, \tilde{A}_{iM} \overset{\text{i.i.d.}}{\sim} \pi_b(\cdot \mid X_i)$$

are independent of (A_i, Y_i), and auxiliary pools are independent across i.

(ii) **Budgets.** The BoN budget N is fixed as $n \to \infty$. For the finite-pool result, $M \geq N - 1$ is fixed. For the efficiency result, $M = M_n \to \infty$.

(iii) **Moments.** For finite-M validity,

$$\mathbb{E}[Y^2] < \infty, \qquad \mathbb{E}[\hat{\mu}_n(X, A)^2] < \infty,$$

where $A \sim \pi_b(\cdot \mid X)$. For the efficiency result, for some $\delta > 0$,

$$\mathbb{E}[|Y|^{2+\delta}] < \infty, \qquad \sup_n \mathbb{E}[|\hat{\mu}_n(X, A)|^{2+\delta}] < \infty.$$

(iv) **reward function consistency.** For the efficiency result,

$$\|\hat{\mu}_n - \mu\|_{L_2(P_{X,A})} = o_p(1),$$

where $P_{X,A}$ is induced by $X \sim P_X$ and $A \sim \pi_b(\cdot \mid X)$.

(v) **Nondegenerate limiting variance.** Define

$$\phi^{\text{eff}}_{N,n}(Z) = \sum_{a \in \mathcal{A}} \pi^{s_n}_N(a \mid X)\mu(X, a) + w^{s_n}_N(X, A)\{Y - \mu(X, A)\} - V(\pi^{s_n}_N).$$

We assume

$$Var(\phi^{\text{eff}}_{N,n}(Z)) \longrightarrow \sigma^2_{N,\text{eff}} \in (0, \infty).$$

The finite-M variances appearing in Theorem 4.2(a) are also assumed positive and finite.

The efficiency statement treats the behavior sampler $\pi_b(\cdot \mid x)$ and, conditional on the training sample, the induced BoN policy as fixed. Thus the unknown components relevant to the efficiency bound are the context distribution and conditional reward distribution.

Unlike generic doubly robust estimators with two learned nuisance functions, no product-rate condition such as

$$\|\hat{\mu} - \mu\|_2 \, \|\hat{w} - w\|_2 = o_p(n^{-1/2})$$

is required. The pooled weight estimator is conditionally unbiased for the exact BoN density ratio, so finite auxiliary sampling introduces mean-zero Monte Carlo variation rather than nuisance-estimation bias. The independent training sample also avoids Donsker or entropy conditions on the fitted model classes.

B.2 Proof of Proposition 4.1

Proof. Fix x and a with $\pi_b(a \mid x) > 0$, and write

$$F^- := F_x^-(a), \qquad F := F_x(a), \qquad q := F - F^-.$$

Here q is the probability that an independent behavior-policy draw has the same score as a, so $q \geq \pi_b(a \mid x) > 0$.

Condition on one of the N BoN candidates being equal to a. If exactly t of the other $N-1$ candidates have equal score and the remaining $N-1-t$ have smaller score, this occurrence of a is selected with probability $1/(t+1)$. By exchangeability of the candidate positions,

$$w_N^s(x,a) = N \sum_{t=0}^{N-1} \binom{N-1}{t} q^t (F^-)^{N-1-t} \frac{1}{t+1} \tag{17}$$

$$= \sum_{k=1}^{N} \binom{N}{k} q^{k-1} (F^-)^{N-k} \tag{18}$$

$$= \frac{F^N - (F^-)^N}{F - F^-}. \tag{19}$$

Using

$$\frac{u^N - v^N}{u - v} = \sum_{\ell=0}^{N-1} u^{N-1-\ell} v^\ell$$

gives

$$w_N^s(x,a) = g_N(F_x(a), F_x^-(a)).$$

This representation also implies

$$0 \leq w_N^s(x,a) \leq N, \tag{20}$$

because g_N is a sum of N terms in $[0,1]$.

Finally,

$$\sum_a \pi_N^s(a \mid x)\hat{\mu}(x,a) = \sum_a \pi_b(a \mid x) w_N^s(x,a)\hat{\mu}(x,a) \tag{21}$$

$$= \mathbb{E}_{A' \sim \pi_b(\cdot|x)} \left[g_N(F_x(A'), F_x^-(A'))\hat{\mu}(x,A') \right], \tag{22}$$

where actions outside the support of π_b have zero probability under BoN. Substituting the same expression for the density ratio in the correction term gives (3). □

B.3 Finite-pool unbiasedness

For a function f, define

$$m_{N,f}(x) := \mathbb{E}_{A \sim \pi_N^s(\cdot|x)}[f(x,A)].$$

Lemma B.2. *For every $M \geq N-1$,*

$$\mathbb{E}[\hat{v}_{i,N}^{s,M+1} \mid X_i] = m_{N,\hat{\mu}}(X_i), \tag{23}$$

$$\mathbb{E}[\hat{w}_{i,N}^{s,M} \mid X_i, A_i] = w_N^s(X_i, A_i). \tag{24}$$

Moreover,

$$0 \leq \hat{w}_{i,N}^{s,M} \leq N.$$

Proof. Conditional on $X_i = x$, the stacked actions

$$A_i, \tilde{A}_{i1}, \ldots, \tilde{A}_{iM}$$

are $M+1$ independent draws from $\pi_b(\cdot \mid x)$. Hence every size-N subset in (5) has the distribution of N fresh BoN candidates. The kernel $h_{i,J}$ analytically averages over uniform tie-breaking, so

$$\mathbb{E}[h_{i,J} \mid X_i = x] = m_{N,\hat{\mu}}(x).$$

Averaging over subsets proves (23).

For completeness, an N-subset has maximum score r_{ig} exactly when it is contained among the E_{ig} observations with score at most r_{ig} but not among the L_{ig} observations with strictly smaller score. There are therefore

$$\binom{E_{ig}}{N} - \binom{L_{ig}}{N}$$

such subsets. Symmetry across the tied maximizing occurrences gives their average prediction $\bar{\mu}_{ig}$, yielding (7).

For the weight estimator, fix $(X_i, A_i) = (x, a)$ and a subset J of $N-1$ auxiliary observations. Define

$$Q_J = N \frac{\mathbf{1}\{s(x, \tilde{A}_{ij}) \le s(x,a) \text{ for all } j \in J\}}{1 + \sum_{j \in J} \mathbf{1}\{s(x, \tilde{A}_{ij}) = s(x,a)\}}.$$

The quantity Q_J/N is exactly the probability that the distinguished logged occurrence a is selected when it is combined with the $N-1$ candidates in J. The argument in Proposition 4.1 therefore gives

$$\mathbb{E}[Q_J \mid X_i = x, A_i = a] = w_N^s(x,a).$$

Averaging over all such subsets proves (24). Grouping subsets according to the number t of equal-score competitors gives exactly (9). Since $0 \le Q_J \le N$, the same bound holds for $\hat{w}_{i,N}^{s,M}$. □

Lemma B.3. *For every $M \ge N-1$,*

$$\mathbb{E}[\psi_N^M(Z_i, \tilde{A}_i)] = V(\pi_N^s),$$

for any fixed reward function satisfying the required moment conditions.

Proof. Lemma B.2 gives

$$\mathbb{E}[\hat{v}_{i,N}^{s,M+1} \mid X_i] = m_{N,\hat{\mu}}(X_i).$$

Also, conditional on (X_i, A_i), the auxiliary pool is independent of Y_i, so

$$\mathbb{E}\left[\hat{w}_{i,N}^{s,M}\{Y_i - \hat{\mu}(X_i, A_i)\} \mid X_i, A_i\right] \tag{25}$$

$$= w_N^s(X_i, A_i)\{\mu(X_i, A_i) - \hat{\mu}(X_i, A_i)\}. \tag{26}$$

Averaging over $A_i \sim \pi_b(\cdot \mid X_i)$ gives

$$m_{N,\mu}(X_i) - m_{N,\hat{\mu}}(X_i).$$

Thus

$$\mathbb{E}[\psi_N^M(Z_i, \tilde{A}_i) \mid X_i] = m_{N,\mu}(X_i),$$

and taking expectation over X_i proves the result. □

B.4 Proof of Theorem 4.2

Proof. For fixed M, conditional on the fitted functions, $(Z_i, \tilde{A}_i)$ are i.i.d. By Lemma B.3,

$$\mathbb{E}[\psi_N^M(Z_i, \tilde{A}_i)] = V(\pi_N^s).$$

The finite-second-moment assumption and the ordinary central limit theorem therefore give

$$\sqrt{n}\left(\hat{V}_{\mathrm{DR}}^M(\pi_N^s) - V(\pi_N^s)\right) \rightsquigarrow \mathcal{N}(0, \sigma_{N,M}^2),$$

proving part (a). No consistency of $\hat{\mu}$ is used.

For part (b), define the exact-ratio DR score

$$\psi_{N,n}^{\infty}(Z) = m_{N,\hat{\mu}_n}(X) + w_N^{s_n}(X,A)\{Y - \hat{\mu}_n(X,A)\}.$$

A fixed-order U-statistic with a square-integrable kernel converges in L_2 to its expectation as the pool size increases; more precisely, its variance is $O(M^{-1})$ for fixed kernel order. Applying this to the plug-in and weight kernels, using $0 \le \hat{w}_N^{s_n,M}, w_N^{s_n} \le N$, yields

$$\left\| \psi_N^{M_n}(Z, \tilde{A}; \hat{\mu}_n) - \psi_{N,n}^{\infty}(Z) \right\|_2 = o_p(1). \tag{27}$$

The moment conditions in Assumption B.1 ensure the corresponding U-statistic second moments are uniformly bounded.

Let

$$\delta_n(x,a) = \hat{\mu}_n(x,a) - \mu(x,a).$$

Since $w_N^{s_n} \le N$,

$$\mathbb{E}[m_{N,\delta_n}(X)^2] \le \mathbb{E}_{X,A\sim\pi_N^{s_n}}[\delta_n(X,A)^2] \tag{28}$$
$$= \mathbb{E}_{X,A\sim\pi_b}[w_N^{s_n}(X,A)\delta_n(X,A)^2] \tag{29}$$
$$\le N\|\delta_n\|_2^2. \tag{30}$$

Hence

$$\left\|\psi_{N,n}^{\infty}(Z) - [m_{N,\mu}(X) + w_N^{s_n}(X,A)\{Y-\mu(X,A)\}]\right\|_2 \tag{31}$$
$$\le (\sqrt{N}+N)\|\hat{\mu}_n - \mu\|_2 = o_p(1). \tag{32}$$

Combining (27) and (32),

$$\left\|\psi_N^{M_n}(Z,\tilde{A};\hat{\mu}_n) - \left[V(\pi_N^{s_n}) + \phi_{N,n}^{\text{eff}}(Z)\right]\right\|_2 = o_p(1).$$

Both scores have expectation $V(\pi_N^{s_n})$. Therefore, if $R_{i,n}$ denotes their difference,

$$\mathbb{E}[R_{i,n}] = 0, \qquad \mathbb{E}[R_{i,n}^2] = o_p(1),$$

and independence across evaluation observations gives

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} R_{i,n} = o_p(1).$$

Consequently,

$$\sqrt{n}\left(\hat{V}_{\text{DR}}^{M_n}(\pi_N^{s_n}) - V(\pi_N^{s_n})\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{N,n}^{\text{eff}}(Z_i) + o_p(1). \tag{33}$$

The $(2+\delta)$ moment condition implies the Lindeberg condition, and the assumed variance convergence therefore yields

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_{N,n}^{\text{eff}}(Z_i) \rightsquigarrow \mathcal{N}(0,\sigma_{N,\text{eff}}^2).$$

Slutsky's theorem proves part (b).

Finally, conditional on the behavior sampler and fitted score, $\pi_N^{s_n}$ is a fixed target policy. The standard AIPW influence function for its value is

$$m_{N,\mu}(X) - V(\pi_N^{s_n}) + w_N^{s_n}(X,A)\{Y-\mu(X,A)\}.$$

Indeed, perturbing the context distribution produces the first term, while perturbing the conditional reward distribution produces the second; because π_b is treated as fixed, there is no additional behavior-policy term. Thus $\phi_{N,n}^{\text{eff}}$ is the efficient influence function in this model, and (33) establishes the claimed efficiency.

The argument also shows why no rate condition between M_n and n is required: the finite-pool approximation is mean zero for every M_n, so it is enough that its per-observation variance vanish as $M_n \to \infty$. □

B.5 Proof of Corollary 5.1

Proof. Write

$$V_N = V(\pi_N^s), \qquad \hat{V}_N = \hat{V}_{\text{DR}}^{M}(\pi_N^s).$$

By Theorem 4.2(a) and finiteness of $\mathcal{N}$,

$$\epsilon_n := \max_{N\in\mathcal{N}} |\hat{V}_N - V_N| = o_p(1).$$

Let $N^\star \in \operatorname{argmax}_{N\in\mathcal{N}} V_N$. Since $\hat{N}_{\text{value}}$ maximizes $\hat{V}_N$,

$$V_{N^\star} \le \hat{V}_{N^\star} + \epsilon_n \le \hat{V}_{\hat{N}_{\text{value}}} + \epsilon_n \le V_{\hat{N}_{\text{value}}} + 2\epsilon_n.$$

Therefore

$$0 \le \max_{N\in\mathcal{N}} V_N - V_{\hat{N}_{\text{value}}} \le 2\epsilon_n = o_p(1),$$

which proves the result. Uniqueness of the maximizing budget is not required because the conclusion concerns policy value rather than the selected index. □

B.6 Proof of Theorem 5.2

Proof. Let

$$\mathcal{N}_+ = \mathcal{N} \setminus \{1\},$$

and collect the improvement scores into

$$D_i = (D_{i,N}^M : N \in \mathcal{N}_+), \qquad \Delta = (\Delta_N : N \in \mathcal{N}_+).$$

By Lemma B.3,

$$\mathbb{E}[D_{i,N}^M] = \Delta_N.$$

Since $\mathcal{N}$ is fixed and finite, the multivariate central limit theorem gives

$$\sqrt{n}(\hat{\Delta}^M - \Delta) \rightsquigarrow \mathcal{N}(0, \Sigma), \qquad \Sigma = Var(D_i),$$

and the empirical covariance matrix converges in probability to Σ.

Conditional on the data,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i (D_i - \hat{\Delta}^M)$$

is exactly Gaussian with the empirical covariance matrix. Hence the multiplier bootstrap consistently estimates the distribution of the maximum standardized estimation error. Therefore

$$\mathbb{P}\left(\Delta_N \ge \widehat{\mathrm{LCB}}_\Delta(N) \text{ for every } N \in \mathcal{N}_+\right) = 1 - \alpha + o(1). \tag{34}$$

On the event in (34), if $\hat{N}_{\mathrm{LCB}} = 1$ then the selected value equals $V(\pi_b)$. Otherwise,

$$\widehat{\mathrm{LCB}}_\Delta(\hat{N}_{\mathrm{LCB}}) > 0,$$

so simultaneous coverage implies

$$\Delta_{\hat{N}_{\mathrm{LCB}}} > 0.$$

Thus

$$\mathbb{P}\left(V(\pi^s_{\hat{N}_{\mathrm{LCB}}}) \ge V(\pi_b)\right) \ge 1 - \alpha - o(1).$$

For oracle consistency, finiteness of the grid gives

$$\max_{N\in\mathcal{N}_+} |\hat{\Delta}_N^M - \Delta_N| = o_p(1).$$

The bootstrap critical value and estimated standard deviations are $O_p(1)$, so the confidence penalties are $o_p(1)$ and therefore

$$\max_{N\in\mathcal{N}_+} \left|\widehat{\mathrm{LCB}}_\Delta(N) - \Delta_N\right| = o_p(1).$$

Let

$$\Delta^\star = \max_{N\in\mathcal{N}_+} \Delta_N.$$

If $\Delta^\star > 0$, the positivity threshold is eventually passed and the finite-grid argmax argument implies that the selected improvement converges to $\Delta^\star$. If $\Delta^\star < 0$, all lower bounds are eventually negative and the procedure selects $N = 1$. If $\Delta^\star = 0$, either the baseline is selected or a budget whose population improvement converges to zero is selected. In all cases,

$$V(\pi^s_{\hat{N}_{\mathrm{LCB}}}) \xrightarrow{p} V(\pi_b) + \max\{0, \Delta^\star\} = \max_{N\in\mathcal{N}} V(\pi^s_N).$$

□

C EXPERIMENTAL DETAILS

C.1 SYNTHETIC EXPERIMENTS

Generative model. Let X be uniform on $\{-1, +1\}$. The action space contains 16 main actions and eight tail actions,

$$\mathcal{A} = \mathcal{A}_{\text{main}} \cup \mathcal{A}_{\text{tail}}.$$

Their base qualities are

$$q_0(a_j) = 0.55 + \frac{0.30(j-1)}{15}, \quad j = 1, \ldots, 16, \qquad q_0(b_k) = 0.05 + \frac{0.20(k-1)}{7}, \quad k = 1, \ldots, 8.$$

The behavior policy is independent of context:

$$\pi_b(a \mid x) = \begin{cases} 0.96/16, & a \in \mathcal{A}_{\text{main}}, \\ 0.04/8, & a \in \mathcal{A}_{\text{tail}}. \end{cases}$$

Conditional rewards satisfy

$$Y \mid X = x, A = a \sim \text{Bernoulli}(\mu(x, a)), \qquad \mu(x, a) = q_0(a) + 0.04x.$$

Independent error controls. We specify the score function and reward function as

$$s_\gamma(x, a) = \mu(x, a) + \gamma \, \mathbf{1}\{a \in \mathcal{A}_{\text{tail}}\},$$
$$\hat{\mu}_\delta(x, a) = \text{clip}(\mu(x, a) + \delta \, \mathbf{1}\{a \in \mathcal{A}_{\text{tail}}\}, 0, 1)\,.$$

Thus, γ changes the ranking of tail actions and hence the target policy $\pi_N^{s_\gamma}$. In contrast, δ changes the reward function predictions while leaving the target policy fixed. At $\delta = 0$, the reward function is correct. The default configuration is $\gamma = 0.65$, $\delta = 0.4$. The selection comparison in Figure 4 uses $\delta = 0.6$, and the efficiency experiment in Appendix D.3 uses $\delta = 0$.

Sampling and evaluation. Each repetition contains n independent logged observations and M independent auxiliary behavior-policy draws per logged context. The main experiments use $n = 4{,}096$, 300 repetitions, and the candidate grid $\mathcal{N} = \{1, 2, 4, \ldots, 256\}$, with $M \geq N_{\max}$, satisfying the shared-pool requirement $M \geq N_{\max} - 1$. Sample-size and auxiliary-pool sweeps vary the corresponding quantity. Shared random streams pair the data across estimators and error settings.

We compute the population policy value exactly:

$$V_N(\gamma) = \frac{1}{2} \sum_{x \in \{-1, +1\}} \sum_{a \in \mathcal{A}} \pi_N^{s_\gamma}(a \mid x) \mu(x, a).$$

The oracle budget is the smallest maximizer of $V_N(\gamma)$ over $\mathcal{N}$. Population values and infinite-pool variance benchmarks are used only for evaluation. The estimators receive sampled observations, score evaluations, and reward function predictions. Experimental regimes were fixed using population quantities before running the repetitions.

C.2 GSM8K ENVIRONMENT

GSM8K provides 7,473 training and 1,319 test questions. We split the training questions deterministically into 3,500 prompts for fitting the reward function and 3,973 validation prompts for off-policy evaluation and selection, and reserve the official test set for the exact evaluation of selected policies. For each prompt we store 512 independently sampled responses with their exact-match rewards and their scores under both reward models. We use $\mathcal{N} = \{1, 2, 4, \ldots, 256\}$ as the candidate grid, with$M \geq N_{\max}$.

Each validation repetition reveals the reward of one uniformly drawn response per prompt; the remaining responses of that prompt form the auxiliary pool used to estimate rank probabilities. This mimics the logged-data setting in which one response per prompt received feedback. We use 500 logging repetitions for estimation and pointwise inference, and 100 repetitions for policy selection.

The reward function is a cubic-spline calibration of the reward model score with six quantile knots, a logistic link, fixed L_2 regularization, and constant extrapolation beyond the observed score range. It

is fit once on the 3,500 training prompts and held fixed across repetitions, budgets, and selection rules. This is a deliberately favorable choice for the Plug-in estimator: the reward function is a calibration of the same score that defines the target policy, so any bias we report is not caused by an unrelated or weak nuisance. Appendix D.7 confirms the conclusion with two independently trained reward functions.

C.3 Baseline estimators

We consider two baseline estimators corresponding to the two components of the pooled BoN-DR estimator in Eq. (11): BoN-IPW and Plug-in. The Plug-in estimator retains only the plug-in term, whereas BoN-IPW uses the estimated BoN density ratio appearing in the correction term.

Pooled BoN-IPW estimator. BoN-IPW uses the estimated BoN density ratio from the correction term of Eq. (11). We define the pooled BoN-IPW estimator as

$$\hat{V}^{M}_{\mathrm{IPW}}(\pi^{s}_{N}) = \frac{1}{n}\sum_{i=1}^{n} \hat{w}^{s,M}_{i,N}(x_i, \bar{a}_i) y_i \tag{35}$$

Thus, BoN-IPW relies on the logged rewards together with the estimated BoN density ratio, without using the plug-in term of BoN-DR.

Plug-in estimator. The Plug-in estimator corresponds directly to the plug-in term of the pooled BoN-DR estimator in Eq. (11). We define the Plug-in estimator as

$$\hat{V}^{M}_{\mathrm{Plug-in}}(\pi^{s}_{N}) = \frac{1}{n}\sum_{i=1}^{n} \hat{v}^{s,M+1}_{i,N}(x_i, \bar{a}_i) \tag{36}$$

The Plug-in estimator therefore evaluates the target BoN policy using the estimated reward function alone, without the correction based on the logged reward. Consequently, errors in the estimated reward function are not corrected by the Plug-in estimator.

C.4 Confidence statements

For DR, $\hat{\sigma}_{N,M}$ is the empirical standard deviation of the scores $\psi^{M}_{N}(z_i, \tilde{a}_i)$; the other estimators use their corresponding scores. Pointwise confidence intervals use the Wald form $\hat{V}^{M}_{\mathrm{DR}}(\pi^{s}_{N}) \pm z_{1-\alpha/2}\hat{\sigma}_{N,M}/\sqrt{n}$. Simultaneous bounds over a candidate grid use the Gaussian multiplier bootstrap of Section 5 with 2,000 draws and $\alpha = 0.05$, with the critical value recomputed for each grid prefix rather than reused from the full grid. Multiplier draws are shared across budgets so that the dependence between value estimates is preserved.

C.5 Metric definitions

Bias is the mean estimate minus the exact value; RMSE is the square root of the mean squared error; the empirical dispersion uses $\mathrm{ddof} = 1$; interval width is the upper minus the lower limit. Coverage is the fraction of repetitions whose interval contains the exact value, counted once per repetition. Simultaneous coverage counts one event per repetition and grid, never one event per budget.

For selection, gain is the exact value of the selected policy minus the exact value at $N = 1$. On GSM8K, estimation and coverage are evaluated against the exact benchmark induced by the stored validation responses, while selected budgets are evaluated independently on the held-out test set. The reported GSM8K harm rate is therefore an empirical analogue of the theoretical no-harm criterion. In the Synthetic experiments, selected policies are evaluated on the exact population. The harm rate is the fraction of repetitions whose gain is negative beyond a numerical tolerance of 10^{-12}; it is not the fraction of individually harmed prompts. Exact ties in a selection objective resolve to the smallest budget. The improvement-LCB rule includes a zero bound at $N = 1$, so it returns $N = 1$ whenever no budget has a strictly positive improvement bound. Monte Carlo uncertainty is binomial for coverage and harm, and the between-repetition standard deviation divided by the square root of the number of repetitions for bias and gain. To keep the tables readable these standard errors are not printed; they are retained for every reported cell in the result files described in Appendix C.6. At 300 repetitions a coverage of 0.95 has a Monte Carlo standard error of 0.013, and at 100 repetitions a harm rate of 0.14 has one of 0.035.

C.6 Reproducibility

All numbers in this paper are computed from a single frozen result set, which is fixed before any figure or table is produced. It contains 5,100 per-repetition records for the Synthetic experiments and 3,600 for GSM8K, covering every combination of environment, estimator, budget, and selection rule reported here and in the appendix. Random number generation uses named streams seeded per repetition, so contexts, logged actions, rewards, and auxiliary draws are shared across estimators and across misspecification settings. Any two estimators are therefore compared on identical data, and differences between them are not attributable to sampling variation in the inputs.

D Additional results for off-policy evaluation

D.1 Reward overoptimization under score function misspecification

We further study how misspecification of the score function affects the induced BoN policy. In the Synthetic experiments, the parameter γ increases the scores of a small set of low-quality actions that are rarely sampled under the behavior policy. As γ increases, these actions are more likely to be selected by BoN, particularly at larger sampling budgets N. Consequently, the induced BoN policy places increasing mass on actions with low probability under the behavior policy, leading to weaker overlap and larger density ratios. Figure 5 illustrates the resulting reward overoptimization: while the expected selected score continues to increase with N, the true policy value can peak and subsequently decline when the score function is sufficiently misspecified. Throughout this experiment, we fix the reward function error at $\delta = 0.4$ and vary only the score-function error γ.

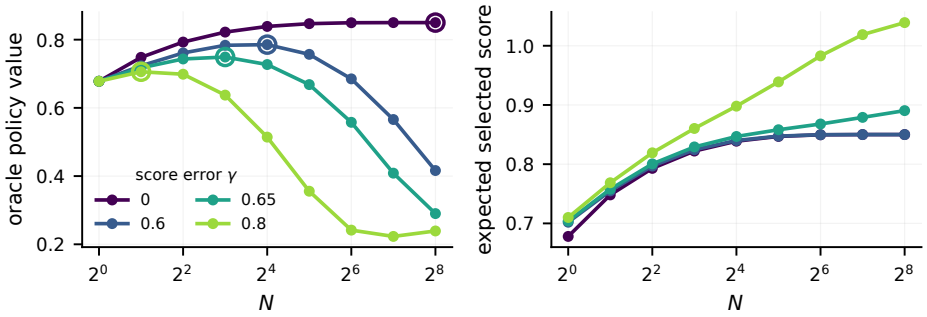


Figure 5: **Robust to score function misspecification** ($\delta = 0.4$). **Left:** exact policy value $V(\pi_N^s)$; circles show the optimal budget $N^\star$. **Right:** expected score of the selected action. Colors show the score error γ.

D.2 Sensitivity to validation-set size

To assess the sensitivity of downstream policy evaluation and budget selection to the amount of validation data, we vary the size of the logged validation set while keeping the remaining experimental setup fixed. Figure 6 reports policy-value estimation at $N = 256$ and value-max selection over the candidate budget grid. As the validation size increases, the RMSE of BoN-DR and BoN-IPW decreases and their confidence-interval coverage approaches the nominal 95% level, whereas the bias of the Plug-in estimator persists. The improved estimation accuracy also translates into more reliable budget selection: the probability of selecting the oracle-optimal budget $N^\star$ increases with the validation size, with BoN-DR reaching high selection accuracy more quickly than BoN-IPW. Overall, the results show that the downstream evaluation and selection procedure becomes increasingly stable as more logged validation data are available.

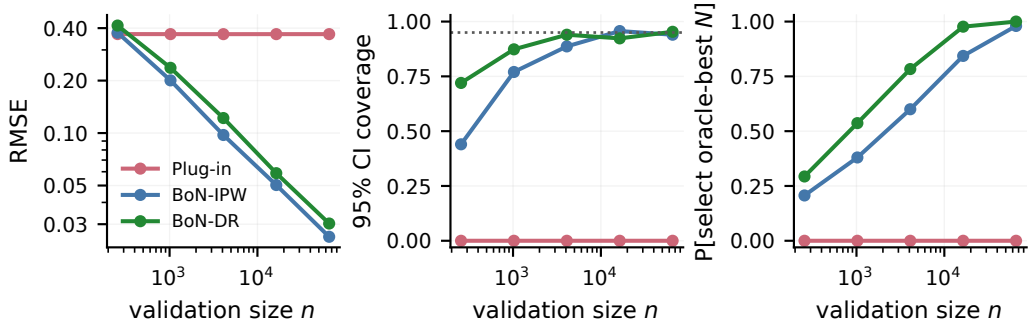


Figure 6: **Estimated BoN policy value at $N = 256$ as logged sample size grows. (a)** RMSE at log scale. **(b)** Coverage of the nominal 95% pointwise confidence interval. The dotted line marks the nominal level. **(c)** Probability of selecting the optimal budget $N^\star$ under value-max selection.

D.3 Finite-M efficiency

Figure 7 isolates the cost of the finite auxiliary pool in the Synthetic experiments with a correctly specified reward function ($\delta = 0$). The paired ratio between the finite-M and the infinite-M DR variance falls from 1.335 at $M = 256$ to 1.019 at $M = 4096$ at $N = 256$, so the Monte Carlo cost of the pool vanishes as M grows, as Theorem 4.2 (b) predicts. The right panel compares the realized variance with the semiparametric efficiency bound: DR attains the bound closely, whereas IPW exceeds it by up to a factor of 4.9.

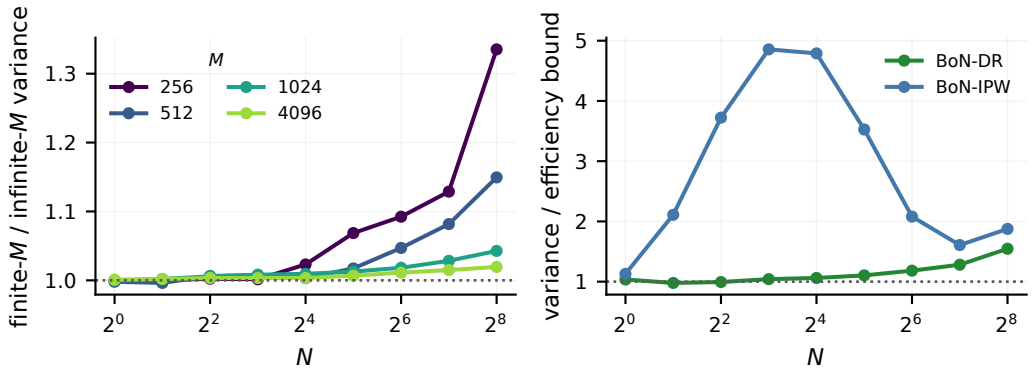


Figure 7: **Finite-M cost and efficiency in the synthetic experiments.** Correctly specified reward function ($\delta = 0$), $\gamma = 0.65$, $n = 4{,}096$, 300 repetitions. **Left:** ratio of the finite-M DR variance to the paired infinite-M oracle variance. **Right:** ratio of the realized variance to the efficiency bound at $M = 256$. The dotted line marks a ratio of one.

D.4 Real coverage

Figure 8 examines interval calibration over the full budget grid on GSM8K. Across both OASST and Skywork, BoN-DR and BoN-IPW achieve coverage close to the nominal 95% level over most budgets, despite the different BoN policies induced by the two score functions. BoN-DR coverage deteriorates at the largest budgets, where rank-weight overlap becomes weak, while BoN-IPW remains better calibrated at these endpoints. This undercoverage is consistent with a finite-sample limitation under increasingly extreme BoN reweighting. In contrast, Plug-in coverage drops to zero from $N = 4$ onward in all six configurations. Overall, the results suggest that BoN-DR and BoN-IPW provide reasonably well-calibrated uncertainty quantification across both score-function regimes, with the main deviations occurring at the largest budgets.

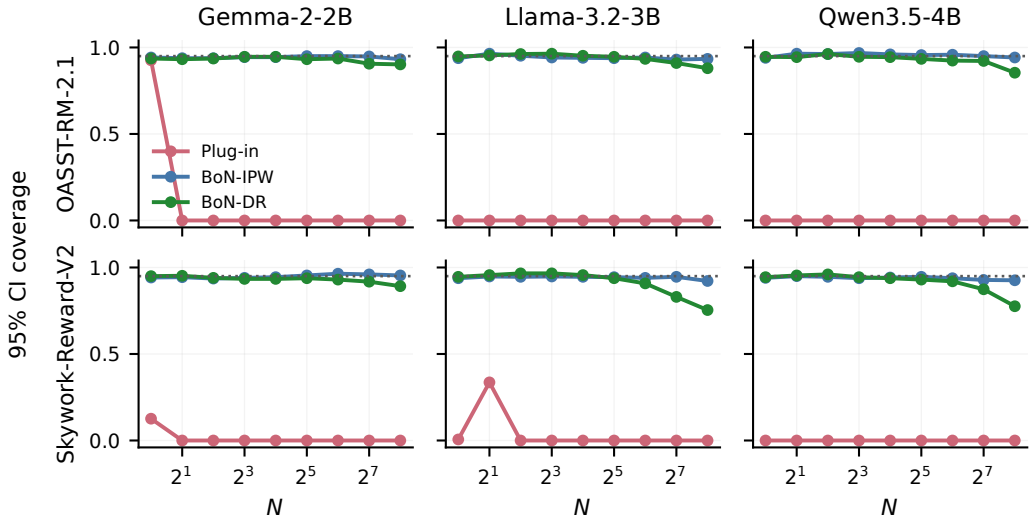


Figure 8: **Interval validity across budgets on GSM8K.** Empirical coverage of the nominal 95% pointwise confidence interval over 500 repetitions. The dotted line marks the nominal level. Rows are external reward models used as score functions, and columns are reference models. The Plug-in estimator never covers from $N = 4$ on. DR declines at the largest budgets, where overlap deteriorates.

D.5 Error decomposition over the budget grid

Figure 9 decomposes the estimation error on GSM8K across all budgets, which a single-budget table cannot show. The Plug-in bias is large and nearly constant in N, whereas DR and IPW are essentially unbiased at every budget and their dispersion grows as overlap deteriorates. The Plug-in estimator therefore becomes competitive in RMSE only at the largest budgets, and only by trading a constant bias against negligible variance. Table 2 gives the per-configuration numbers at $N = 256$.

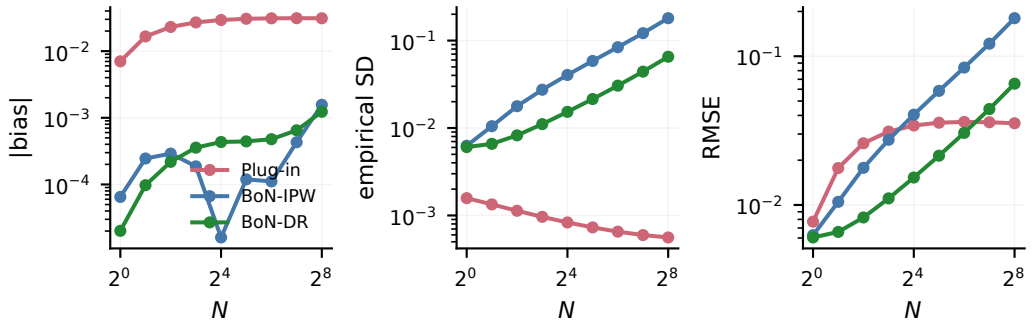


Figure 9: **Absolute bias, dispersion, and RMSE on GSM8K over the budget grid.** Each quantity is averaged over the six configurations, 500 repetitions, log scale.

Table 2: **Estimation on GSM8K at $N = 256$.** 500 repetitions with $n = 3{,}973$ logged prompts. Coverage is that of the nominal 95% pointwise confidence interval. **Bold** marks the lowest RMSE and the coverage closest to 0.95 within each configuration.

Reward model	Estimator	Bias	RMSE	CI width	Coverage
OASST-RM-2.1, reference model Gemma-2-2B					
	Plug-in	0.097	**0.097**	0.002	0.000
	BoN-IPW	0.002	0.156	0.625	**0.932**
	BoN-DR	0.008	0.102	0.364	0.902
OASST-RM-2.1, reference model Llama-3.2-3B					
	Plug-in	0.029	**0.029**	0.001	0.000
	BoN-IPW	0.005	0.179	0.703	**0.934**
	BoN-DR	0.003	0.079	0.280	0.880
OASST-RM-2.1, reference model Qwen3.5-4B					
	Plug-in	0.018	**0.018**	0.001	0.000
	BoN-IPW	0.008	0.187	0.726	**0.942**
	BoN-DR	0.003	0.058	0.225	0.854
Skywork-Reward-V2, reference model Gemma-2-2B					
	Plug-in	0.041	**0.041**	0.007	0.000
	BoN-IPW	0.004	0.172	0.704	**0.954**
	BoN-DR	-0.003	0.064	0.252	0.892
Skywork-Reward-V2, reference model Llama-3.2-3B					
	Plug-in	-0.013	**0.013**	0.002	0.000
	BoN-IPW	-0.010	0.198	0.737	**0.922**
	BoN-DR	-0.002	0.040	0.134	0.754
Skywork-Reward-V2, reference model Qwen3.5-4B					
	Plug-in	0.015	**0.015**	0.001	0.000
	BoN-IPW	0.001	0.190	0.727	**0.926**
	BoN-DR	-0.002	0.048	0.168	0.776

D.6 Overlap diagnostics

Figure 10 reports the effective sample size of the rank weights on GSM8K. It falls from 3,973 prompts at $N = 1$ to about 25 at $N = 256$ and about 10 at $N = 512$, where only 0.5% of prompts carry non-zero weight. The endpoint $N = 512$ uses all $M + 1 = 512$ stored responses and satisfies the boundary condition $M = N - 1$ in Theorem 4.2. We report it only as an overlap stress test and restrict the main text to $N \leq 256$. Extending the grid to $N_{\max} = 512$ roughly doubles the residual harm of DR under improvement-LCB selection, from 0.12 to 0.24 on Llama with OASST and from 0.14 to 0.33 on Qwen with OASST, while IPW is unaffected at 0.00 and 0.01.

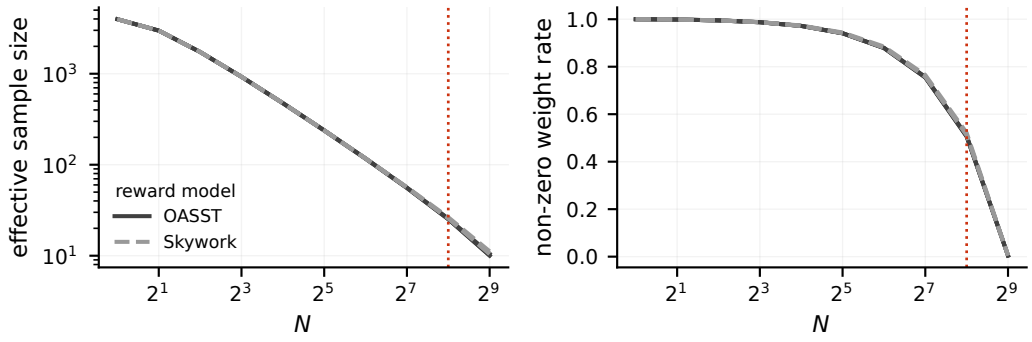


Figure 10: **Overlap of the rank weights on GSM8K. Left:** effective sample size, log scale, out of 3,973 validation prompts. **Right:** fraction of prompts receiving non-zero weight. Solid lines are OASST, dashed lines are Skywork; the vertical line marks the largest budget used in the main text.

D.7 Robustness to the reward function

Table 3 repeats the GSM8K comparison with two reward functions trained independently of the score function, a TF–IDF logistic regression and an embedding-based LightGBM model. Plug-in coverage is zero for both, so the failure reported in RQ2 is a property of the estimator under misspecification and not an artifact of the score-calibration spline used in the main text.

Table 3: **Estimation on GSM8K at $N = 256$ with independently trained reward functions.** OASST score function. Coverage is that of the nominal 95% pointwise confidence interval. **Bold** marks the lowest RMSE within each reference model and reward function.

Reward estimator	Estimator	Bias	RMSE	Coverage
Reference model Gemma-2-2B				
Embedding LightGBM	Plug-in	0.040	**0.040**	0.000
	BoN-IPW	0.008	0.167	0.880
	BoN-DR	-0.015	0.093	0.970
TF–IDF logistic	Plug-in	0.022	**0.022**	0.000
	BoN-IPW	0.008	0.167	0.880
	BoN-DR	-0.011	0.101	0.940
Reference model Llama-3.2-3B				
Embedding LightGBM	Plug-in	0.023	**0.023**	0.000
	BoN-IPW	-0.011	0.194	0.920
	BoN-DR	0.005	0.090	0.800
TF–IDF logistic	Plug-in	0.024	**0.024**	0.000
	BoN-IPW	-0.011	0.194	0.920
	BoN-DR	0.006	0.084	0.820
Reference model Qwen3.5-4B				
Embedding LightGBM	Plug-in	0.018	**0.018**	0.000
	BoN-IPW	0.013	0.185	0.940
	BoN-DR	0.003	0.061	0.830
TF–IDF logistic	Plug-in	0.014	**0.014**	0.000
	BoN-IPW	0.013	0.185	0.940
	BoN-DR	0.002	0.058	0.830

E Additional results for selection

Table 4 gives the numerical results and median selected budgets for Figure 4. The following subsections report the complete comparisons across configurations.

Table 4: **Selecting N under challenging conditions.** *Synthetic experiments:* score error $\gamma = 0.65$, reward function error $\delta = 0.6$, $n = 4{,}096$, 300 repetitions; gain in policy value. *GSM8K:* Qwen3.5-4B with OASST, $n = 3{,}973$, 100 logging repetitions; gain in accuracy points. Gain is measured against $N = 1$ and higher is better; harm is the fraction of repetitions with negative gain and lower is better; $\widetilde{N}$ is the median selected budget. $N_{\max} = 256$. **Bold** marks the best value per column among the selection methods. The last row gives the best attainable gain on the grid.

	Synthetic exp (misspecified)			GSM8K: Qwen3.5-4B / OASST		
Selection method	Gain	Harm	$\widetilde{N}$	Gain (pp)	Harm	$\widetilde{N}$
Naive BoN	-0.388	1.00	256	-1.50	1.00	256
Plug-in	-0.380	1.00	256	-0.28	1.00	4
BoN-IPW, value-max	0.068	**0.00**	8	-0.84	0.77	32
BoN-IPW, improvement-LCB	0.062	**0.00**	4	**0.00**	**0.01**	1
BoN-DR, value-max	**0.069**	**0.00**	8	-0.82	0.73	64
BoN-DR, improvement-LCB	0.067	**0.00**	4	-0.19	0.14	1
Oracle-best budget	*0.071*	*0.00*	*8*	*0.00*	*0.00*	*1*

E.1 IMPROVEMENT-LCB REDUCES HARMFUL BUDGET CHOICES.

Figure 4 expands the selection analysis in Figure 11 by reporting both the achieved gain and the harm rate in two challenging regimes. The Synthetic experiments with $\gamma = 0.65$ and $\delta = 0.6$ has an interior optimal budget, $N^\star = 8$, so successful selection must exploit the benefit of increasing N without overoptimizing at larger budgets. In contrast, for Qwen3.5-4B with OASST on GSM8K, $N^\star = 1$, so any departure from the reference policy is harmful. These two regimes therefore separate the ability to capture a beneficial nontrivial budget from the ability to avoid unnecessary increases in inference-time compute.

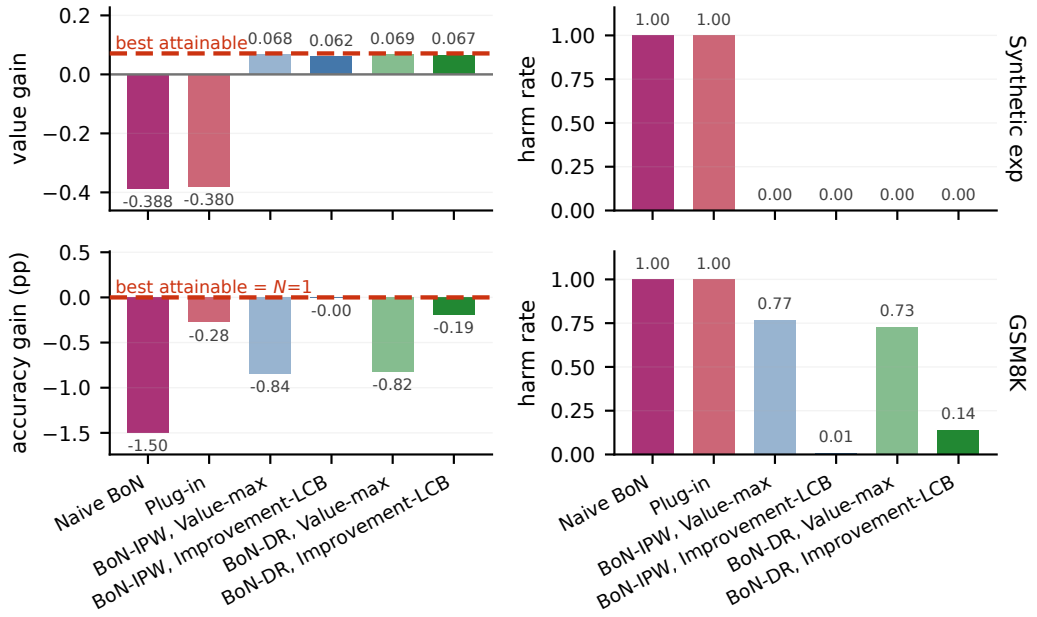


Figure 11: **Improvement-LCB reduces harmful budget choices. Top:** Synthetic experiments with $\gamma = 0.65$, $\delta = 0.6$, $n = 4{,}096$, and 300 repetitions. **Bottom:** Qwen3.5-4B with OASST on GSM8K, with 3,973 validation prompts and 100 logging repetitions. **Left:** mean gain over $N = 1$, measured in policy value for the Synthetic experiments and accuracy percentage points for GSM8K. Higher is better. The dashed red line marks the best attainable gain on the grid. **Right:** harm rate. Lower is better. Light and solid bars show value-max and improvement-LCB, respectively. Plug-in uses value-max, and Naive BoN selects $N_{\max} = 256$. Labels report numerical values.

In the Synthetic experiments, value-max selection with BoN-IPW and BoN-DR recovers nearly all of the attainable gain and incurs no observed harm, whereas Naive BoN and Plug-in are harmful in every repetition. Improvement-LCB is slightly more conservative, but retains most of the gain while likewise avoiding harmful selections. Thus, when the policy-value estimates are sufficiently accurate, both direct value maximization and uncertainty-aware selection identify useful nontrivial budgets.

The GSM8K regime is more challenging because the reference policy itself is optimal, so the best attainable gain over $N = 1$ is zero. Value-max selection nevertheless frequently chooses larger, harmful budgets: BoN-IPW and BoN-DR lose 0.84 and 0.82 accuracy percentage points on average, with harm rates of 77% and 73%, respectively. Improvement-LCB substantially reduces these losses and lowers the harm rates to 1% for BoN-IPW and 14% for BoN-DR. The remaining BoN-DR harm is consistent with its finite-sample undercoverage at the largest budgets discussed in Appendix D.4. Thus, accounting for uncertainty is particularly useful when maximization over noisy value estimates would otherwise favor unnecessarily large budgets.

E.2 Complete results on GSM8K

Table 5 reports every estimator with value-max and improvement-LCB for all six configurations, including the aligned ones omitted from RQ3. Two patterns hold throughout. Under the aligned Skywork reward model, no rule incurs harm, and improvement-LCB reduces gain without reducing the observed harm rate, most visibly for IPW. Under OASST, improvement-LCB reduces the harm incurred by value-max selection with DR and IPW. Residual harm remains for DR on Llama and Qwen.

Table 5: **Complete selection results on GSM8K.** $N_{\max} = 256$, 100 logging repetitions. Gain is measured in accuracy points against $N = 1$ and higher is better; harm is the fraction of repetitions with negative gain and lower is better. Naive BoN does not depend on the selection rule and is reported once per configuration.

Naive BoN			Plug-in		BoN-IPW		BoN-DR	
Gain	Harm	Rule	Gain	Harm	Gain	Harm	Gain	Harm
OASST-RM-2.1, reference model Gemma-2-2B								
4.23	0.00	Value-max	4.84	0.00	4.45	0.00	4.51	0.00
		Improvement-LCB	4.84	0.00	3.25	0.00	3.98	0.00
OASST-RM-2.1, reference model Llama-3.2-3B								
-0.51	1.00	Value-max	0.33	0.00	-0.03	0.44	-0.09	0.53
		Improvement-LCB	0.33	0.00	0.01	0.00	-0.05	0.12
OASST-RM-2.1, reference model Qwen3.5-4B								
-1.50	1.00	Value-max	-0.28	1.00	-0.84	0.77	-0.82	0.73
		Improvement-LCB	-0.28	1.00	0.00	0.01	-0.19	0.14
Skywork-Reward-V2, reference model Gemma-2-2B								
23.71	0.00	Value-max	23.71	0.00	21.86	0.00	22.91	0.00
		Improvement-LCB	23.71	0.00	16.17	0.00	20.57	0.00
Skywork-Reward-V2, reference model Llama-3.2-3B								
15.92	0.00	Value-max	15.92	0.00	13.95	0.00	15.07	0.00
		Improvement-LCB	15.92	0.00	8.86	0.00	14.12	0.00
Skywork-Reward-V2, reference model Qwen3.5-4B								
3.11	0.00	Value-max	3.11	0.00	2.57	0.00	2.91	0.00
		Improvement-LCB	3.11	0.00	0.38	0.00	2.23	0.00

E.3 Complete results in the Synthetic experiments

Table 6 sweeps the reward function error at a fixed overoptimized score function. It shows why RQ1 reports $\delta = 0.6$: all estimator-based selection rules have zero observed harm up to $\delta = 0.4$, because the Plug-in bias is not yet large enough to move its selected budget past the point at which the policy value falls below the value at $N = 1$. At $\delta = 0.6$ the Plug-in budget jumps to $N_{\max}$ and harms in every repetition, while DR and IPW retain positive gains with zero observed harm across the tested values of δ. This also shows that a biased estimator can still select a safe budget, which is why estimation quality and selection quality are evaluated separately in RQ1.

Table 6: **Complete selection results in the Synthetic experiments.** Score error $\gamma = 0.65$, $n = 4{,}096$, $N_{\max} = 256$, 300 repetitions. Gain is measured in policy value against $N = 1$ and higher is better; harm is the fraction of repetitions with negative gain and lower is better. Naive BoN does not depend on the selection rule and is reported once per value of δ.

Naive BoN			Plug-in		BoN-IPW		BoN-DR	
Gain	Harm	Rule	Gain	Harm	Gain	Harm	Gain	Harm
Reward estimator error $\delta = -0.2$								
-0.388	1.00	Value-max	0.065	0.00	0.068	0.00	0.070	0.00
		Improvement-LCB	0.065	0.00	0.062	0.00	0.067	0.00
Reward estimator error $\delta = 0$								
-0.388	1.00	Value-max	0.071	0.00	0.068	0.00	0.070	0.00
		Improvement-LCB	0.071	0.00	0.062	0.00	0.067	0.00
Reward estimator error $\delta = 0.2$								
-0.388	1.00	Value-max	0.071	0.00	0.068	0.00	0.070	0.00
		Improvement-LCB	0.071	0.00	0.062	0.00	0.067	0.00
Reward estimator error $\delta = 0.4$								
-0.388	1.00	Value-max	0.049	0.00	0.068	0.00	0.070	0.00
		Improvement-LCB	0.049	0.00	0.062	0.00	0.067	0.00
Reward estimator error $\delta = 0.6$								
-0.388	1.00	Value-max	-0.380	1.00	0.068	0.00	0.069	0.00
		Improvement-LCB	-0.319	1.00	0.062	0.00	0.067	0.00

E.4 Selection across the range of score error

Figure 12 varies the score error at a fixed reward function. Naive BoN is optimal while the score function is aligned and collapses as soon as it is exploitable, whereas every estimator-based rule has nonnegative mean gain across the tested range. DR attains the lowest regret against the optimal budget.

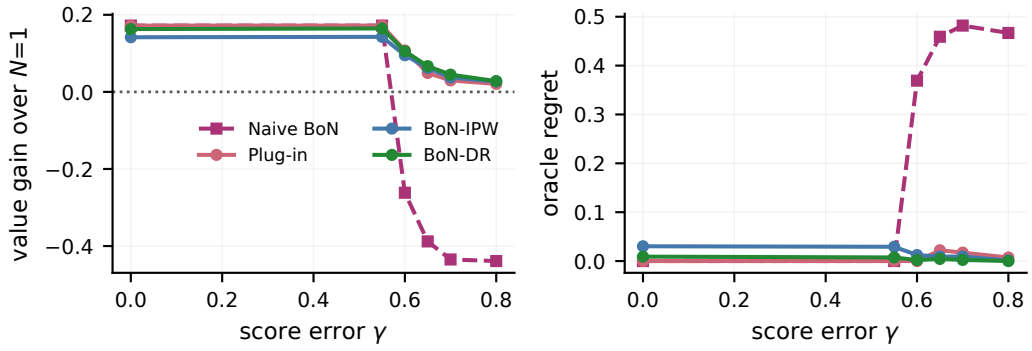


Figure 12: **Selection across the range of score error in the Synthetic experiments.** Improvement-LCB rule, $\delta = 0.4$, $n = 4{,}096$, $N_{\max} = 256$, 300 repetitions. **Left:** gain over $N = 1$. **Right:** regret against the optimal budget on the grid, lower is better.

F Extension to preference rewards

Our BoN evaluation framework extends to settings in which the scalar reward encodes a pairwise preference between two model responses, as in AI alignment and RLHF. For a prompt or context X, let

$$A = (A_1, A_2)$$

denote a pair of candidate responses and let

$$Y \in \{0, 1\}$$

be the observed reward indicating whether the first response is preferred to the second response. Thus, $Y = 1$ means that A_1 is preferred to A_2.

We assume that the behavior comparison policy factorizes as

$$\pi_b(a_1, a_2 \mid x) = \pi_{b,1}(a_1 \mid x)\pi_{b,2}(a_2 \mid x).$$

Here, $\pi_{b,1}$ denotes the policy from which the first candidate response is sampled, and $\pi_{b,2}$ denotes the policy from which the second, reference candidate response is sampled. In an RLHF application, $\pi_{b,1}$ could be the base, supervised fine-tuned, or previous model from which BoN samples are drawn. The policy $\pi_{b,2}$ is the opponent or reference policy used in the comparison. Often, the two policies are the same, so that

$$\pi_{b,1} = \pi_{b,2} = \pi_{\mathrm{ref}},$$

corresponding to the common design in which both candidate responses are sampled independently from the same reference model. More generally, however, the two marginals may differ, for example if the first response is sampled from a current model and the second response from a fixed baseline model.

Let $s(x, a)$ denote the score function used for BoN selection. Given a prompt x, the BoN policy induced by $\pi_{b,1}$ samples

$$\widetilde{A}_1, \ldots, \widetilde{A}_N \overset{\text{iid}}{\sim} \pi_{b,1}(\cdot \mid x)$$

and returns

$$J_N(x) := \operatorname*{argmax}_{1 \le j \le N} s(x, \widetilde{A}_j),$$

$$I_N(x) \mid \widetilde{A}_1, \ldots, \widetilde{A}_N \sim \mathrm{Uniform}(J_N(x)), \qquad A_N^\star := \widetilde{A}_{I_N(x)}.$$

Ties are thus broken uniformly among maximizing sampled occurrences, including repeated responses. We denote the resulting policy by $q_N(\cdot \mid x)$.

F.1 Pairwise Win-Rate Objective

In the preference-learning setting, a natural target estimand is the pairwise win rate of the BoN policy against a reference policy ρ:

$$W_N(\rho) = \mathbb{E}_X \mathbb{E}_{A \sim q_N(\cdot|X),\, B \sim \rho(\cdot|X)} \left[\mu(X, A, B)\right],$$

where

$$\mu(x, a, b) = \mathbb{P}(Y = 1 \mid X = x, A_1 = a, A_2 = b)$$

is the conditional mean reward function, here the probability that response a is preferred to response b.

The reference policy ρ specifies the opponent against which the BoN policy is evaluated. A particularly convenient and practically relevant choice is

$$\rho = \pi_{b,2}.$$

In this case, the target asks how often a BoN response sampled from the first behavior marginal beats a response sampled from the same reference distribution that generated the second response in the logged preference data. When

$$\pi_{b,1} = \pi_{b,2} = \rho = \pi_{\mathrm{ref}},$$

the estimand is the probability that a BoN response beats an ordinary response sampled from the same reference model.

A common modeling assumption in preference learning is the Bradley–Terry model, under which there exists a latent reward function $r_0(x, a)$ such that

$$\mu(x, a, b) = \sigma\{r_0(x, a) - r_0(x, b)\},$$

where $\sigma(u) = (1 + \exp(-u))^{-1}$ is the logistic function. Given a sample $\{(X_i, A_{i1}, A_{i2}, Y_i)\}_{i=1}^n$, a reward model r_θ can be estimated by minimizing the negative log-likelihood

$$\mathcal{L}_{\mathrm{BT}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\Big[Y_i \log \sigma\{r_\theta(X_i, A_{i1}) - r_\theta(X_i, A_{i2})\} + (1 - Y_i)\log \sigma\{r_\theta(X_i, A_{i2}) - r_\theta(X_i, A_{i1})\}\Big].$$

The corresponding reward function is

$$\widehat{\mu}(x, a, b) = \sigma\{\widehat{r}(x, a) - \widehat{r}(x, b)\}.$$

As in Section 4.3, fitted functions are constructed independently of the evaluation data and auxiliary draws. The Bradley–Terry model is useful because it connects pairwise preferences to the reward modeling perspective common in RLHF. However, the pairwise win-rate estimand does not require the Bradley–Terry model for identification. It is identified as long as the conditional preference regression $\mu(x, a, b)$ is identified on the relevant support.

F.2 Doubly Robust Estimation

The target pair distribution for the win-rate estimand is

$$q_N^{\text{pair}}(a, b \mid x) = q_N(a \mid x)\rho(b \mid x),$$

whereas the behavior pair distribution is

$$\pi_b(a, b \mid x) = \pi_{b,1}(a \mid x)\pi_{b,2}(b \mid x).$$

Therefore, the pairwise density ratio is

$$w_N(x, a, b) = \frac{q_N(a \mid x)\rho(b \mid x)}{\pi_{b,1}(a \mid x)\pi_{b,2}(b \mid x)}.$$

Define

$$m_N(x; \rho) = \mathbb{E}_{A\sim q_N(\cdot|x),\, B\sim\rho(\cdot|x)}\left[\mu(x, A, B)\right].$$

Then the efficient influence function for $W_N(\rho)$ in the nonparametric model for μ is

$$\phi_N(O) = m_N(X; \rho) - W_N(\rho) + w_N(X, A_1, A_2)\{Y - \mu(X, A_1, A_2)\}.$$

This yields the doubly robust estimator

$$\widehat{W}_N(\rho) = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{m}_N(X_i; \rho) + \widehat{w}_N(X_i, A_{i1}, A_{i2})\{Y_i - \widehat{\mu}(X_i, A_{i1}, A_{i2})\}\right].$$

The plug-in component can be estimated by Monte Carlo:

$$\widehat{m}_N(x; \rho) = \frac{1}{M}\sum_{m=1}^{M}\widehat{\mu}(x, \widetilde{A}^{\star}_{m,N}, \widetilde{B}_m),$$

where

$$\widetilde{A}^{\star}_{m,N} = \widetilde{A}_{m,I_{m,N}}, \qquad \widetilde{A}_{mj} \overset{\text{iid}}{\sim} \pi_{b,1}(\cdot \mid x), \qquad \widetilde{B}_m \sim \rho(\cdot \mid x).$$

Here, conditional on the sampled candidates, $I_{m,N}$ is drawn uniformly from $\operatorname{argmax}_{1\le j\le N} s(x, \widetilde{A}_{mj})$. The M terms above are independent Monte Carlo replicates, each using N candidate draws and an independent opponent draw; this is a direct Monte Carlo construction rather than the pooled estimator of Section 4.2.

F.3 BoN density ratios

The density ratio can be simplified using the structure of the BoN policy. Suppose q_N is induced by sampling N candidates independently from $\pi_{b,1}(\cdot \mid x)$ and selecting the response with the largest score $s(x, a)$. Define

$$F^{-}_{x,1}(a) = \mathbb{P}_{A\sim\pi_{b,1}(\cdot|x)}\{s(x, A) < s(x, a)\},$$
$$F_{x,1}(a) = \mathbb{P}_{A\sim\pi_{b,1}(\cdot|x)}\{s(x, A) \le s(x, a)\}.$$

Proposition 4.1 gives, for $\pi_{b,1}(a \mid x) > 0$,

$$\frac{q_N(a \mid x)}{\pi_{b,1}(a \mid x)} = g_N\big(F_{x,1}(a), F^{-}_{x,1}(a)\big).$$

Thus,

$$w_N(x,a,b) = g_N\big(F_{x,1}(a), F^-_{x,1}(a)\big) \cdot \frac{\rho(b \mid x)}{\pi_{b,2}(b \mid x)}.$$

The strict and weak score probabilities can be estimated by Monte Carlo sampling from the first behavior marginal. Given samples

$$\bar{A}_1, \ldots, \bar{A}_L \overset{\text{iid}}{\sim} \pi_{b,1}(\cdot \mid x),$$

we estimate

$$\widehat{F}^-_{x,1}(a) = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}\{s(x,\bar{A}_\ell) < s(x,a)\},$$

$$\widehat{F}_{x,1}(a) = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}\{s(x,\bar{A}_\ell) \le s(x,a)\}.$$

This gives the plug-in estimate of the BoN ratio

$$\widehat{R}_N(x,a) = g_N\Big(\widehat{F}_{x,1}(a), \widehat{F}^-_{x,1}(a)\Big).$$

This plug-in ratio estimate is generally biased at finite L. The unbiased alternative is Eq. (9), applied to the first behavior marginal with $M = L \ge N-1$; the plug-in ratio estimate does not inherit its finite-pool unbiasedness. The resulting pairwise weight estimate is

$$\widehat{w}_N(x,a,b) = \widehat{R}_N(x,a) \cdot \frac{\rho(b \mid x)}{\pi_{b,2}(b \mid x)}.$$

A particularly important special case is

$$\rho = \pi_{b,2}.$$

Then the second density ratio cancels and

$$w_N(x,a,b) = \frac{q_N(a \mid x)}{\pi_{b,1}(a \mid x)} = g_N\big(F_{x,1}(a), F^-_{x,1}(a)\big).$$

Hence the doubly robust estimator becomes

$$\widehat{W}_N(\pi_{b,2}) = \frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{m}_N(X_i;\pi_{b,2}) + \widehat{R}_N(X_i,A_{i1})\{Y_i - \widehat{\mu}(X_i,A_{i1},A_{i2})\}\Big].$$

If, additionally,

$$\pi_{b,1} = \pi_{b,2} = \rho = \pi_{\text{ref}},$$

then both logged responses are independent samples from the same reference model, the BoN policy is obtained by reranking samples from that same model, and the target value is

$$W_N(\pi_{\text{ref}}) = \mathbb{E}_{X,A\sim q_N(\cdot|X),\, B\sim\pi_{\text{ref}}(\cdot|X)}\left[\mu(X,A,B)\right].$$

In words, $W_N(\pi_{\text{ref}})$ is the probability that a BoN response beats a standard response from the reference model. In this common case, the only density ratio required for debiasing is the BoN ratio

$$R_N(x,a) = g_N\big(F_{x,1}(a), F^-_{x,1}(a)\big),$$

which can be estimated using samples from the reference model alone.

This extension shows that the proposed BoN evaluation approach carries over naturally to preference-learning settings. The key requirement is that the logged comparison policy factorizes, and that the first behavior marginal is the policy from which BoN candidates are sampled. If the behavior pair distribution does not factorize, then the required density ratio generally depends on the full joint distribution of (A_1, A_2) and cannot be recovered from samples from the marginal policies alone.