PRIVACY WITHOUT REGRET: DIFFERENTIALLY PRIVATE INFERENCE-TIME ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-of-N (BoN) sampling is the simplest and most widely deployed inference-time alignment strategy, but it suffers from two distinct problems: reward hacking, in which the selected response exploits errors in the proxy reward model, and the absence of any privacy protection for the sensitive human preference data used to train that reward model. We show that a single intervention – adding calibrated noise to reward scores before selection – resolves both. Our first result, Private Best-of-N (PrivBoN), establishes that Gumbel noise at an appropriate scale simultaneously provides ϵ-differential privacy and implements KL-regularized alignment. Whenever the privacy budget exceeds a critical threshold ϵ^*, the privacy-mandated noise is the regret-optimal regularization, and privacy imposes zero additional alignment cost—matching the information-theoretic skyline of Huang et al. (2025). Because ϵ^* depends on an unknown coverage coefficient, we introduce Private Inference-Time Pessimism (PrivITP), which combines χ^2-regularized rejection sampling with a two-phase Gaussian mechanism. PrivITP achieves ex-post (ϵ, δ)-DP with a privacy cost independent of the number of responses n, cleanly decouples the regularization parameter from the privacy parameter, and attains the skyline up to a noise-inflation term. Experiments across several language models, datasets, and reward models confirm our results: PrivBoN and PrivITP are scaling-monotonic (unlike BoN, which degrades past a critical n), and PrivITP outperforms PrivBoN at equivalent privacy levels, with the largest gains in the strong-privacy regime.

1 INTRODUCTION

Aligning large language models (LLMs) with human preferences is a central challenge in modern AI (Stiennon et al., 2020). While training-time methods such as RLHF (Christiano et al., 2017) and DPO (Rafailov et al., 2023) have received extensive attention, a parallel line of work focuses on *inference-time alignment*: improving the quality of a frozen model's outputs at serving time, without updating its parameters. Inference-time alignment has become increasingly important as frontier models are deployed via APIs to millions of users, making it impractical to fine-tune per-user or per-application preferences into the model weights. The simplest and most widely deployed inference-time strategy is Best-of-N (BoN) sampling (Nakano et al., 2021; Gui et al., 2024), which generates n candidate responses from a base policy π_0, scores each with a learned reward model $\hat{r}$, and returns the top-scoring candidate. BoN is attractive for its simplicity – it requires no gradient computation, works with any frozen model, and scales with test-time compute (Snell et al., 2024).

A growing body of recent work has established tight theoretical characterizations of BoN's KL vs. win-rate tradeoff (Beirami et al., 2025; Yang et al., 2024) and proposed smoothed variants that interpolate between hard maximization and random selection (Verdun et al., 2025; Aminian et al.; Khalaf et al.). However, BoN is known to be vulnerable to *reward hacking* (Gao et al., 2023; Khalaf et al.): as the number of responses n grows, BoN increasingly selects responses in which the learned reward $\hat{r}$ overestimates the true reward r^*, thereby degrading alignment. Huang et al. (2025) formalizes this by showing that BoN is not scaling-monotonic – more candidates can yield *worse* alignment and that BoN's regret crucially depends on the quality of the learned reward model $\hat{r}$.

Reward models are typically trained on sensitive human preference data collected from crowd workers or domain experts whose judgments encode personal values and cultural context. Privacy attacks on such data are well-documented, spanning membership inference (Feng et al., 2024), preference

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.

poisoning (Carlini et al., 2021; Wu et al., 2025; Baumgärtner et al., 2024; Wang et al., 2024), and downstream-output leakage (Tramèr et al., 2022). This risk is diluted at frontier scale, where aggregation over millions of annotations shrinks per-example influence, but sharp where aggregation is weak, e.g., domain-expert or personalized RMs (tens to hundreds of credentialed annotators). These are also the settings where per-example sensitivity is non-negligible and inference-time alignment is most attractive. Best-of-N selection is directly exposed. Consider a reward model $\hat{r}_{\mathcal{D}}$ deployed as a BoN service: an adversary wishing to determine whether a preference example (x_0, y_0^+, y_0^-) was in $\mathcal{D}$ can craft prompts whose plausible completions are structured around y_0^+ and y_0^-, query repeatedly, and observe shifts in the empirical distribution of returned responses. Since BoN's argmax is deterministic, even small training-attributable shifts in rewards manifest as measurable shifts in selection frequencies, and over a deployment of T queries the cumulative leakage is uncontrolled.

While differential privacy has been widely applied to training-time alignment (SFT, reward-model training, and RLHF) (Yu et al., 2025; Wang et al., 2025), inference-time privacy remains largely unstudied. We fill this gap by introducing *differentially private inference-time alignment*: mechanisms ensuring that the selected response satisfies ε-DP with respect to D, bounding the information any adversary can extract about any individual preference—regardless of query budget or side information. We make the following *contributions.*

(1) Private Best-of-N (PrivBoN, Section 3). We show that adding i.i.d. Gumbel noise with scale $\sigma = 2\Delta_r/\varepsilon$ to reward scores before BoN selection yields a mechanism that simultaneously satisfies ε-DP and implements the KL-regularized alignment optimum in the large-n limit, where Δ_r is the reward-model sensitivity on D. When the privacy budget exceeds the critical threshold $\varepsilon^*(x) = 2\Delta_r\sqrt{C^{\pi^*}(x)}/\varepsilon_{\mathrm{RM}}(x)$, the privacy-mandated noise coincides with the regret-optimal regularization and privacy becomes "free"—PrivBoN's regret $O(\sqrt{C^{\pi^*}\varepsilon_{\mathrm{RM}}^2})$ matches the information-theoretic skyline of Huang et al. (2025) at zero additional alignment cost. Here $\varepsilon_{\mathrm{RM}}$ denotes reward-model error and C^{π^*} the coverage coefficient of a comparator policy π^*.

(2) Private Inference-Time Pessimism (PrivITP, Section 4). PrivBoN's "privacy is free" threshold depends on the unknown coverage coefficient C^{π^*}, making it unverifiable in practice. Our main contribution, PrivITP, resolves this by combining χ^2-regularized rejection sampling with a two-phase Gaussian mechanism. We prove an ex-post approximate-DP guarantee with an n-independent bound that depends only on the realized halting time, via a *reformulation lemma* that maps ReLU-based rejection sampling onto a randomized-threshold mechanism. The regret bound $O(\sqrt{C^{\pi^*}(\varepsilon_{\mathrm{RM}}^2 + R_{\max}\sigma)})$ cleanly separates alignment error from privacy cost and matches the skyline up to noise inflation, where $R_{\max}$ is the maximum reward; the analysis relies on a *smoothed-weight policy comparison lemma* that bounds value gaps via the ℓ_∞-distance of policy weight functions. The ex-post structure further enables adaptive composition via the FSRC framework (Lebensold et al., 2024), answering substantially more queries under a fixed privacy budget than standard composition.

(3) Empirical validation (Section 5). Across four reward models and two language models of varying sizes on GSM8K, MMLU, and MATH datasets, we demonstrate that the theoretical results hold: Private algorithms eliminate BoN's reward-hacking pathology while preserving its scaling benefits; PrivITP consistently matches or outperforms PrivBoN at matched privacy levels; and the ex-post composition structure translates into substantial multi-query deployment gains.

PrivITP and PrivBoN are, to our knowledge, the first inference-time alignment algorithms whose regret matches the information-theoretic skyline under formal differential privacy guarantees. We view inference-time DP as complementary to training-time DP: the latter amortizes its privacy cost across all future queries but permanently degrades the model, whereas ours leaves the model intact and spends a per-query budget. Neither dominates – inference-time DP fits settings where the reward model is not under deployer's control (e.g. a frozen checkpoint) or where degradation is unacceptable, while training-time DP is preferable at high query volume where its one-time cost amortizes.

2 PRELIMINARIES

Inference-time alignment. We consider a setting in which a user submits a prompt $x \in \mathcal{X}$ and the system must return a single response $y \in \mathcal{Y}$. The system has access to a base language model (the *reference policy*) $\pi_0 : \mathcal{X} \to \Delta(\mathcal{Y})$, assumed to have full support on $\mathcal{Y}$, and a reward model $\hat{r}_D : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ trained on a dataset D. The true reward function $r^* : \mathcal{X} \times \mathcal{Y} \to [0, R_{\max}]$ is

unknown, and the trained reward model $\hat{r}_D$ serves as its proxy. The base policy π_0 is a trained language model (e.g., an instruction-tuned model using SFT and/or RLHF), the true reward r^* captures the desired objective (e.g., alignment with human preferences or correctness under a proof checker), and proxy reward $\hat{r}_D$ is a custom-trained model on a (possibly sensitive) dataset D.

Given a base policy π_0, reward model $\hat{r}_D$, and a prompt $x \in \mathcal{X}$, the goal of the algorithm designer is to construct a policy $\hat{\pi}$, which generates high-quality responses as measured by the true reward $r^*(x, y)$. This is formalized using the notion of *inference-time regret* (Huang et al., 2025). The *regret* of a policy $\hat{\pi}$ relative to a comparator policy π^* is defined as

$$\mathrm{Reg}(\hat{\pi}; x) = J(\pi^*; x) - J(\hat{\pi}; x), \tag{1}$$

where $J(\pi; x) = \mathbb{E}_{y\sim\pi(\cdot|x)}[r^*(x, y)]$ denotes the value of a policy π for a given prompt x. Without assumptions on the accuracy of the reward model $\hat{r}_D$ and on the coverage of the base policy π_0, achieving low regret is not possible (Huang et al., 2025). To this end, define the reward model error $\varepsilon(x, y) = r^*(x, y) - \hat{r}_D(x, y)$. The expected squared error under the base policy,

$$\varepsilon^2_{\mathrm{RM}}(x) = \mathbb{E}_{y\sim\pi_0(\cdot|x)}[\varepsilon(x, y)^2] \,,$$

measures the mean-squared discrepancy between proxy and rewards. We abstract away reward model training by assuming access to a model $\hat{r}_D$ with squared-error $\varepsilon^2_{\mathrm{RM}}(x)$ and study how regret depends on this error and what algorithmic interventions can make it small. The χ^2*-coverage coefficient*

$$C^\pi(x) = \mathbb{E}_{y\sim\pi(\cdot|x)}\left[\tfrac{\pi(y|x)}{\pi_0(y|x)}\right] = 1 + 2\,\chi^2(\pi(\cdot|x)\|\pi_0(\cdot|x))$$

measures how well the base policy π_0 covers a given policy π, where $\chi^2(p\|q)$ denotes the χ^2-divergence between probability distributions p, q. A comparator policy π^* with a larger C^{π^*} indicates a harder alignment target. Empirical studies have shown that standard language models provide sufficient coverage of high-quality responses, enabling performance gains at inference time. Huang et al. (2025) lays down the importance of coverage by establishing a regret skyline in the *sample and evaluate* framework, where, for a given prompt x, the learner can sample n responses $\{y_j\}_j$ from the base policy π_0, and observe their likelihood $\pi_0(y_j|x)$ and reward scores $\hat{r}_D(x, y_j)$.

Proposition 2.1 (Regret skyline (Huang et al., 2025))**.** *Fix a prompt x and base policy π_0. For any inference-time alignment algorithm $\hat{\pi}$, there exist a reward function r^* and reward model $\hat{r}_D$ with error $\varepsilon_{\mathrm{RM}}(x)$ such that* $\mathrm{Reg}(\pi; x) \geq \frac{1}{4}\sqrt{C^{\pi^*}\varepsilon^2_{\mathrm{RM}}(x)}$.

This lower bound is tight: Huang et al. (2025) achieves it with InferenceTimePessimism, a χ^2-regularized rejection sampling algorithm. A central question of this work is whether differential privacy – which requires adding noise that ostensibly degrades performance – can be achieved without exceeding this skyline. To this end, we first formalize the privacy definitions.

Differential privacy. We protect the training data D of the reward model. Two datasets D and D' from a database $\mathcal{D}$ are *adjacent*, written $D \sim D'$, if they differ in a single individual's data. The *sensitivity* of the reward model is $\Delta_r = \sup_{x,y} \sup_{D\sim D'} |\hat{r}_D(x, y) - \hat{r}_{D'}(x, y)|$, which quantifies the maximum influence of any one training point on any reward score.

Definition 2.2. *A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{S}$ satisfies:*

- *ε-DP if* $\Pr[\mathcal{M}(D) = s] \leq e^\varepsilon \Pr[\mathcal{M}(D') = s]$ *for all $s \in \mathcal{S}$ and all $D \sim D'$.*
- *(ε, δ)-DP if* $\Pr[\mathcal{M}(D) = s] \leq e^\varepsilon \Pr[\mathcal{M}(D') = s] + \delta$ *for all $s \in \mathcal{S}$ and all $D \sim D'$.*
- *ε_p-ex-post DP if $\exists$ a function $\varepsilon_p : \mathcal{D} \to \mathcal{S}$ such that* $\Pr[\mathcal{M}(D) = s] \leq e^{\varepsilon_p(s)} \Pr[\mathcal{M}(D') = s]$ *for all $s \in \mathcal{S}$ and all $D \sim D'$.*

Traditional ex-ante privacy (Dwork & Roth, 2014) bounds the guarantee worst-case over outcomes, before the mechanism's randomness is realized; ex-post privacy (Lebensold et al., 2024) bounds it conditional on the realized output, allowing queries with favorable outcomes to "save" budget for later queries under a fixed total.

3 PrivBoN: Private Best-of-N

We introduce the PrivBoN (Private Best-of-N) framework, which adds calibrated noise to reward scores before selection. The mechanism is simple: sample n candidates $y_1, \ldots, y_n$ inde-

Algorithm 1 PrivBoN: Private Best-of-N

Require: Prompt x, base policy π_0, reward model $\hat{r}_D$, number of candidates n, noise scale $\sigma > 0$
1: **for** $i = 1$ to n **do**
2: Sample candidate $y_i \sim \pi_0(\cdot|x)$ and noise $g_i \sim \mathrm{Gumbel}(0, \sigma)$
3: Compute noisy reward $\tilde{r}_i \leftarrow \hat{r}_D(x, y_i) + g_i$
4: **end for**
5: **return** y_{i^*}, where $i^* = \arg\max_i \tilde{r}_i$

pendently from $\pi_0(\cdot|x)$, perturb each reward score $\hat{r}_D(x, y_i)$ with independent noise g_i, and return the candidate with the highest noisy reward. The noise distribution determines the privacy guarantee and the induced policy. We use the $\mathrm{Gumbel}(0, \sigma)$ distribution, which has the PDF $f(x) = \frac{1}{\sigma}\exp(-x/\sigma - \exp(-x/\sigma))$ and the CDF $F(x) = \exp(-\exp(-x/\sigma))$.

Theorem 3.1 (PrivBoN privacy)**.** *Let $\hat{r}_D(x, y)$ be a reward model trained on a dataset D with sensitivity Δ_r. Then, Algorithm 1 with $\sigma = 2\Delta_r/\varepsilon$ satisfies ε-DP.*

The result follows from a classical connection: adding Gumbel noise and taking the argmax is equivalent to sampling from the softmax distribution (thanks to the Gumbel-max trick), which is precisely the exponential mechanism of McSherry and Talwar (McSherry & Talwar, 2007). It also yields a clean, closed-form expression for the induced policy, given below.

Proposition 3.2 (PrivBoN policy)**.** *PrivBoN with Gumbel noise induces the policy*

$$\pi_{\mathrm{PrivBoN}}(y|x) = n\,\pi_0(y|x)\,\mathbb{E}_{y_{2:n}\sim\pi_0(\cdot|x)}\left[\frac{e^{\hat{r}_D(x,y)/\sigma}}{\sum_{j=1}^{n} e^{\hat{r}_D(x,y_j)/\sigma}}\right],$$

where $y_1 = y$ is fixed and $y_{2:n}$ are drawn independently from π_0. Assuming the reward distribution under π_0 is continuous, (a) it recovers the standard BoN policy as $\sigma \to 0$ and (b) it converges pointwise to the KL-regularized optimum $\pi_\infty(y|x) \propto \pi_0(y|x)\exp(\hat{r}_D(x, y)/\sigma)$ as $n \to \infty$.

Part (b) is a crucial property: PrivBoN implements, in the large-n limit, the KL-regularized optimum $\max_\pi \mathbb{E}_{y\sim\pi}[\hat{r}_D(x, y)] - \sigma\mathrm{KL}(\pi\|\pi_0)$. The noise scale σ simultaneously controls privacy budget and regularization strength – raising the question of whether the two impose compatible demands on σ.

3.1 Regret guarantee

We now analyze how the noise required for privacy affects alignment quality within the regret framework of Huang et al. (2025). Our main result is that privacy-mandated noise implements the regularization needed for optimal alignment whenever the privacy budget exceeds a threshold ε^*.

Theorem 3.3 (PrivBoN regret)**.** *Let $r^*(x, y) \in [0, R_{\max}]$ be the true reward and $\hat{r}_D(x, y)$ be a reward model trained on a dataset D. Let the reward model error $\epsilon(x, y) \sim \mathcal{N}(0, \sigma_r^2(x))$ under the base policy π_0. Let π^* be any comparator policy with finite coverage $C^{\pi^*}(x) < \infty$. Then, for $n = \widetilde{\Omega}\left(e^{(R_{\max}+\varepsilon_{\mathrm{RM}})/\sigma}\log(R_{\max}/\varepsilon_{\mathrm{RM}})\right)$, the regret of PrivBoN satisfies*

$$\mathrm{Reg}(\pi_{\mathrm{PrivBoN}}; x) \lesssim \underbrace{\sigma C^{\pi^*}(x)}_{\text{KL-bias}} + \underbrace{\varepsilon_{\mathrm{RM}}^2(x)/\sigma}_{\text{overoptimization}} + \underbrace{\sqrt{C^{\pi^*}(x)}\,\varepsilon_{\mathrm{RM}}(x)}_{\text{irreducible}} + \underbrace{\frac{R_{\max}e^{\frac{R_{\max}+\varepsilon_{\mathrm{RM}}\sqrt{\log n}}{\sigma}}}{\sqrt{n}}}_{\text{finite sample}}. \quad (2)$$

where $\varepsilon_{\mathrm{RM}}^2(x) = \mathbb{E}_{y\sim\pi_0(\cdot|x)}[\epsilon(x, y)^2] = \sigma_r^2(x)$ is the reward model variance.

Optimal noise scale. The third term is the skyline. The first two terms exhibit a bias-variance tradeoff in σ: a large σ (strong regularization, strong privacy) incurs bias from under-exploiting the reward signal, whereas a small σ (weak regularization, weak privacy) incurs overoptimization from exploiting reward model errors. Balancing them yields $\sigma^*(x) = \varepsilon_{\mathrm{RM}}(x)/\sqrt{C^{\pi^*}(x)}$, at which the regret $\mathrm{Reg}(\pi_{\mathrm{PrivBoN}}; x) \le 3\sqrt{C^{\pi^*}(x)\varepsilon_{\mathrm{RM}}^2(x)}$ matches the information-theoretic skyline (Proposition 2.1) in the large-n limit.

"Privacy is free" threshold. When σ is set by the privacy budget ($\sigma = 2\Delta_r/\varepsilon$), the regret becomes

$$\mathrm{Reg}(\pi_{\mathrm{PrivBoN}}; x) \le \frac{2\Delta_r C^{\pi^*}(x)}{\varepsilon} + \frac{\varepsilon\cdot\varepsilon_{\mathrm{RM}}^2(x)}{2\Delta_r} + \sqrt{C^{\pi^*}(x)}\varepsilon_{\mathrm{RM}}(x). \quad (3)$$

The privacy-mandated noise σ is no larger than the regret-optimal noise $\sigma^*(x)$ when

$$\varepsilon \geq \varepsilon^*(x) = \frac{2\Delta_r\sqrt{C^{\pi^*}(x)}}{\varepsilon_{\mathrm{RM}}(x)}. \tag{4}$$

For all $\varepsilon \geq \varepsilon^*(x)$, the regret matches the skyline: the noise required for privacy is no more than what alignment already needs. E.g., for a reward model with $\varepsilon_{\mathrm{RM}} = 0.35, \Delta_r = 0.1, C^{\pi^*} = 2$, this gives $\varepsilon^* \approx 0.8$, i.e., any privacy budget above 0.8 is free. Below $\epsilon < \varepsilon^*(x)$, the first term in (3) dominates and the regret degrades gracefully as $O\big(\frac{\Delta_r C^{\pi^*}(x)}{\varepsilon} + \sqrt{C^{\pi^*}(x)}\varepsilon_{\mathrm{RM}}(x)\big)$. The Gaussian-error assumption yields the closed-form threshold $\epsilon^*(x)$; under a general MSE assumption, the qualitative conclusions persist, but the threshold loses its closed form. See Appendix B.4.

The regret bound (2) goes down with n for $n \geq n_0(\sigma, R_{\max}, \epsilon_{\mathrm{RM}})$, confirming that *more candidates never worsen the upper bound*, which rules out the catastrophic reward-hacking regime of BoN where the bound degrades with an increase in n. The finite-sample term, however, carries an $R_{\max}e^{(R_{\max}+\epsilon_{\mathrm{RM}}\sqrt{\log n})/\sigma}/\sqrt{n}$ prefactor: for typical values ($R_{\max} = 5$, $\sigma = 1$) the asymptotic rate kicks in only for $n \gtrsim 10^4$, and for moderate n the regret behaves as $O(R_{\max}e^{R_{\max}/\sigma}/\sqrt{n})$. PrivITP (Section 4) avoids this exponential dependence via rejection sampling.

Remark 3.4 (Privacy-mandated noise as optimal regularization)**.** *Both privacy and alignment require softening the* $\arg\max$ *to prevent overfitting to the proxy reward, and both implement this softening through the same mechanism – Gumbel noise of scale σ, equivalent via the Gumbel-max trick (McSherry & Talwar, 2007) to softmax with temperature σ. Privacy imposes $\sigma \geq 2\Delta_r/\varepsilon$, alignment requires $\sigma^*(x) \leq \varepsilon_{\mathrm{RM}}(x)/\sqrt{C^{\pi^*}(x)}$. The two are compatible whenever $\varepsilon \geq \varepsilon^*(x)$.*

Remark 3.5 (Comparison with prior work)**.** *SoftBoN (Aminian et al.; Khalaf et al.) is algorithmically equivalent to PrivBoN but analyzed differently: Aminian et al. bound the gap within the SoftBoN policy class at a fixed temperature, whereas we bound regret against an arbitrary comparator π^* and match the skyline. Khalaf et al. treat temperature as a free tuning parameter, whereas we derive it from the privacy budget, revealing the threshold $\varepsilon^*(x)$ at which privacy becomes free.*

Remark 3.6 (Inference-time vs. training-time DP)**.** *A natural alternative could be to train the reward model once with ε-DP and run non-private selection. However, DP-training degrades the RM, raising its error to $\varepsilon_{\mathrm{RM}}^{\mathrm{train}}(\varepsilon) > \varepsilon_{\mathrm{RM}}$ for every $\varepsilon > 0$, and post-processing makes this* permanent*: every free pays the raised irreducible floor $\sqrt{C^{\pi^\star}}\,\varepsilon_{\mathrm{RM}}^{\mathrm{train}}(\varepsilon)$ (Proposition 2.1). Our mechanism operates on the undegraded RM and, above $\varepsilon^\star$, pays the smaller $\sqrt{C^{\pi^\star}}\,\varepsilon_{\mathrm{RM}}$ per query.*

3.2 Reward hacking robustness

Theorem 3.3 establishes that PrivBoN achieves low regret. A separate and complementary question is whether PrivBoN prevents *reward hacking*—the pathology where the selected response exploits errors in the proxy reward, yielding high $\hat{r}_D$ but low r^* (Gao et al., 2023; Khalaf et al.). We quantify this through two measures: the win-rate against the base policy and the hacking gap $\mathbb{E}_{\hat{\pi}}[\epsilon(x, y)]$, which directly measures the extent to which the selected response overfits to reward-model errors. For a given prompt x, the win-rate (Gui et al., 2024) of a policy $\hat{\pi}$ against π_0 is

$$p_{\hat{\pi}\succ\pi_0|x} = \Pr_{Y\sim\hat{\pi},\, Y_0\sim\pi_0}\big(\hat{r}_D(x, Y) \geq \hat{r}_D(x, Y_0)\big).$$

High win-rate does not imply low regret when $\hat{r}_D$ and r^* are mismatched: BoN achieves win-rate $n/(n+1)$ (Beirami et al., 2025) while simultaneously incurring regret $\widetilde{\Omega}(\sqrt{\varepsilon_{\mathrm{RM}}^2 n})$ (Huang et al., 2025). The two quantities diverge because BoN's argmax overfits to $\hat{r}_D$'s errors. The next results shows that PrivBoN's softmax selection caps both quantities at finite limits, eliminating this pathology.

Proposition 3.7 (Win-rate saturation and bounded hacking)**.** *Let $\epsilon(x, y) \sim \mathcal{N}(0, \sigma_r^2(x))$ under π_0. As $n \to \infty$, we have (a) BoN's win-rate converges to 1, whereas PrivBoN's saturates at $\Phi(\sigma_r(x)/(\sigma\sqrt{2})) < 1$ and (b) BoN's hacking gap grows without bound as $O(\sigma_r(x)\sqrt{\log n})$, while PrivBoN's converges to the finite limit $\sigma_r^2(x)/\sigma$.*

The finite saturation limits reflect the same mechanism: the privacy/regularization noise σ prevents the selection from fully exploiting the reward signal, even with unlimited candidates. Consequently, there exists a crossover $n^*(\sigma, \sigma_r)$ beyond which PrivBoN achieves strictly higher *true* reward than

Algorithm 2 Private InferenceTimePessimism (PrivITP)

Require: Prompt x, base policy π_0, reward model $\hat{r}_D$, candidates n, regularization $\beta > 0$, noise scales $\sigma_X, \sigma_Z > 0$, truncation parameter L

Phase 1 (private normalization):
  Draw $y_1, \ldots, y_n \sim \pi_0(\cdot|x)$; compute $r_i = \hat{r}_D(x, y_i)$
  Find λ such that $\frac{1}{n}\sum_{i=1}^{n} \mathrm{relu}\big(\beta^{-1}(r_i - \lambda)\big) = 1$
  Set $\tilde{\lambda} = \lambda + \mathcal{N}(0, \sigma_X^2)$ and $M = (R_{\max} + \sigma_Z L - \tilde{\lambda})/\beta$
Phase 2 (private rejection sampling):
for $i = 1$ to n **do**
  Draw $y_i' \sim \pi_0(\cdot|x)$; Compute $\tilde{r}_i = \hat{r}_D(x, y_i') + \mathcal{N}(0, \sigma_Z^2)$
  Set $w_i = \mathrm{relu}\big(\beta^{-1}(\tilde{r}_i - \tilde{\lambda})\big)$; Sample $\xi_i \sim \mathrm{Ber}(\min(w_i/M, 1))$
  if $\xi_i = 1$ **then**
    return y_i'
  end if
end for
return $y_{n+1}' \sim \pi_0(\cdot|x)$

BoN – BoN's expected true reward eventually decreases with n (Goodhart's law), while PrivBoN's converges monotonically to $\mathbb{E}_{\pi_\infty}[r^*]$. For typical reward models (e.g., Oasst on GSM8K dataset), n^* falls well below 100 (see Figure 1(top row)). Proofs are deferred to Appendix B.

4 Private Inference Time Pessimism

PrivBoN's threshold $\varepsilon^*(x)$ depends on the unknown coverage coefficient $C^{\pi^*}(x)$ (4), so a practitioner cannot verify whether their privacy budget falls in the free regime. Private InferenceTimePessimism (PrivITP) resolves this by combining the χ^2-regularized rejection sampling of Huang et al. (2025) with calibrated noise injection, decoupling reward-hacking mitigation (controlled by the regularization) from privacy (controlled by the noise).

Algorithm. PrivITP operates in two phases, each consuming independent noise. Phase 1 draws n candidates from $\pi_0(\cdot|x)$ and privately releases the normalization constant $\tilde{\lambda}$ of the χ^2-regularized policy by adding Gaussian noise to the solution λ of the empirical normalization equation (we omit dependence on x for brevity). Phase 2 draws a fresh batch of n candidates, perturbs their rewards with Gaussian noise, and performs rejection sampling with an acceptance threshold $\tilde{\lambda}$ from Phase 1. Independence of the two batches is essential for both the privacy and regret analyses; pseudocode appears in Algorithm 2. The ReLU in the acceptance weights implements pessimism: candidates whose noisy reward falls below $\tilde{\lambda}$ receive zero weight, discarding responses that the reward model does not confidently rank above the population average. This mechanism prevents reward hacking independently of the injected noise, so privacy and alignment can be tuned separately (Theorem 4.3).

Privacy guarantee. Each phase accesses the dataset D through reward evaluations. The injected Gaussian noise in the normalizer in Phase 1 and in the rewards in Phase 2 ensures that the released quantities ($\tilde{\lambda}$ from Phase 1 and the selected response from Phase 2) are differentially private.

Theorem 4.1 (PrivITP privacy)**.** *Let $\hat{r}_D(x, \cdot) \in [0, R_{\max}]$ with sensitivity Δ_r. Algorithm 2 with Phase 1 scalar Gaussian noise σ_X and Phase 2 query noise σ_Z satisfies $(\varepsilon_1 + \varepsilon_2^{\mathrm{post}}(t),\ \delta)$-DP* ex-post*, where **Phase 1** contributes (ε_1, δ)-DP with $\varepsilon_1 = \Delta_r\sqrt{2\ln(1.25/\delta)}/\sigma_X$ and **Phase 2** contributes* pure *ex-post $\varepsilon_2^{\mathrm{post}}(t)$-DP, where t is the halting time (index of the first accepted candidate):*

$$\varepsilon_2^{\mathrm{post}}(t) = (t-1)\,\log \frac{\mathbb{E}_{u\sim\mathrm{Unif}[0,1]}\Big[\Phi\Big(\frac{\tilde{\lambda}+\beta M u - R_{\max}+\Delta_r}{\sigma_Z}\Big)\Big]}{\mathbb{E}_{u\sim\mathrm{Unif}[0,1]}\Big[\Phi\Big(\frac{\tilde{\lambda}+\beta M u - R_{\max}}{\sigma_Z}\Big)\Big]} + \log \frac{\mathbb{E}_{u\sim\mathrm{Unif}[0,1]}\Big[\Phi\Big(\frac{\Delta_r - \tilde{\lambda} - \beta M u}{\sigma_Z}\Big)\Big]}{\mathbb{E}_{u\sim\mathrm{Unif}[0,1]}\Big[\Phi\Big(\frac{-\tilde{\lambda} - \beta M u}{\sigma_Z}\Big)\Big]}.$$

Unlike standard privacy composition over n candidates (which gives a worst-case bound proportional to n), Theorem 4.1 gives a bound, which is independent of n, computable via one-dimensional numerical integration and proportional to the *actual halting time* t. The coefficient of the linear term,

$$\kappa := \log \frac{\mathbb{E}_u[\Phi((\tilde{\lambda} + \beta M u - R_{\max} + \Delta_r)/\sigma_Z)]}{\mathbb{E}_u[\Phi((\tilde{\lambda} + \beta M u - R_{\max})/\sigma_Z)]},$$

is the per-round privacy cost of rejection, which depends on the relative positions of $\tilde{\lambda}$, $R_{\max}$, and σ_Z. When σ_Z is large relative to Δ_r (strong privacy), κ is small; when σ_Z is comparable to Δ_r, κ dominates the privacy budget. The bound requires the reward model's outputs to lie in a bounded range $[0, R_{\max}]$, which is a mild assumption in practice as reward models for language alignment are typically trained with bounded output heads (e.g., sigmoid or clipped outputs) Huang et al. (2025).

Remark 4.2 (Ex-post vs. ex-ante accounting)**.** *Theorem 4.1 gives an ex-post guarantee: the privacy cost depends on the realized halting time t. For worst-case accounting, $\epsilon_2^{\text{post}}(n)$ upper-bounds the per-query cost. In practice, rejection happens quickly when the reward model is informative ($\mathbb{E}[t] = O(1)$), so the ex-post bound is substantially tighter and enables the composition of Section 4.1.*

Regret guarantee. We now analyze the regret of PrivITP against an arbitrary comparator policy π^*. The central result is that PrivITP achieves the information-theoretic skyline (Proposition 2.1)with a noise inflation and it does so while satisfying pure ex-post ε-DP (Theorem 4.1).

Theorem 4.3 (PrivITP regret)**.** *Let $r^*(x, y) \in [0, R_{\max}]$ be the true (unknown) reward and $\hat{r}_D \colon \mathcal{X} \times \mathcal{Y} \to [0, R_{\max}]$ be a reward model trained on a dataset D with mean-squared error $\varepsilon^2_{\text{RM}}(x) = \mathbb{E}_{y\sim\pi_0(x)}[(\hat{r}_D(x, y) - r^*(x, y))^2]$. Let π^* be any comparator policy with finite coverage $C^{\pi^*}(x) < \infty$. For regularization $\beta \geq 2\sigma_X$, truncation parameter $L = O(\log n)$ and $n = \max\left\{\Omega\Big(\frac{R_{\max}}{\beta}\log\Big(\frac{R_{\max}}{\beta\,\epsilon_{\text{RM}}}\Big)\Big), \widetilde{\Omega}\left(\frac{R_{\max}+\sigma_Z}{\beta}\right)\right\}$, PrivITP (Algorithm 2) satisfies*

$$\text{Reg}(\pi_{\text{PrivITP}}; x) \lesssim \underbrace{\beta \cdot C^{\pi^*}(x)}_{\chi^2\ \textit{bias}} + \underbrace{\beta^{-1}(\varepsilon^2_{\text{RM}}(x) + R_{\max}\sigma)}_{\textit{overoptimization+noise}} + \underbrace{\sqrt{C^{\pi^*}(x)}\varepsilon_{\text{RM}}(x)}_{\textit{irreducible}} + \underbrace{R_{\max}/\sqrt{n}}_{\textit{finite sample}}, \quad (5)$$

where $\sigma = \sigma_X + \sigma_Z$ denotes the total noise level.

As with PrivBoN, the bound in (5) is non-increasing in n for $n \geq n_0(\beta, \sigma_Z, R_{\max}, \epsilon_{\text{RM}})$, ruling out the reward-hacking regime where adding candidates provably degrades performance.

Optimal regularization. The first two terms exhibit a bias-variance tradeoff in β: a large σ (strong regularization) incurs bias from under-exploiting the reward signal, whereas a small β (weak regularization) incurs overoptimization from exploiting reward model errors. Balancing the first two terms yields $\beta^*(x) = \sqrt{(\varepsilon^2_{\text{RM}}(x) + R_{\max}\sigma)/C^{\pi^*}(x)}$, at which the regret,

$$\text{Reg}(\pi_{\text{PrivITP}}; x) \leq 3\sqrt{C^{\pi^*}(x)\big(\varepsilon^2_{\text{RM}}(x) + R_{\max}\sigma\big)}\,,$$

matches the information-theoretic skyline (proposition 2.1) in the large-n limit (up to the noise inflation due to privacy). Proofs are deferred to Appendix C.

Remark 4.4 (Compatibility of $\beta^* \geq 2\sigma_X$)**.** *The condition $\beta \geq 2\sigma_X$ in Theorem 4.3 is satisfied by $\beta^*(x)$ whenever $\sigma_X \leq \frac{1}{2}\sqrt{(\epsilon^2_{\text{RM}} + R_{\max}\sigma)/C^{\pi^*}(x)}$, a mild bound translating to a lower bound on ϵ_1 in the privacy-calibrated regime. When violated, setting $\beta = 2\sigma_X \geq \beta^*(x)$ recovers the bound with an additive $2\sigma_X C^{\pi^*}(x)$ penalty in the χ^2-bias term.*

Privacy-mandated noise. The Phase 2 query noise σ_Z governs the pure ex-post pure-DP guarantee of Theorem 4.1: smaller σ_Z yields larger per-query privacy cost $\varepsilon_2^{\text{post}}(t)$. To quantify the trade-off, we use the order-of-magnitude relation obtained by linearizing (4.1) in the small-Δ_r/σ_Z regime: the per-round log-ratio is of order Δ_r/σ_Z, giving $\varepsilon_2^{\text{post}}(t) \asymp t \cdot \Delta_r/\sigma_Z$, or equivalently $\sigma_Z \asymp t \cdot \Delta_r/\varepsilon_2^{\text{post}}(t)$. For Phase 1 per-query privacy cost (ε_1, δ), we set noise $\sigma_X \asymp \Delta_r\sqrt{\ln(1/\delta)}/\varepsilon_1$. Substituting into (4) yields the privacy-mandated regret bound

$$\text{Reg}(\pi_{\text{PrivITP}}; x) \leq 3\sqrt{C^{\pi^*}(x)\left(\varepsilon^2_{\text{RM}}(x) + \frac{R_{\max} \cdot t\Delta_r}{\varepsilon_2^{\text{post}}(t)} + \frac{R_{\max}\Delta_r\sqrt{\ln(1/\delta)}}{\epsilon_1}\right)}. \quad (6)$$

On the relationship to prior work. PrivITP combines the χ^2-regularized rejection sampling of Huang et al. (2025) with the Gaussian above-threshold mechanism of Lebensold et al. (2024). The technical novelty lies in the analysis of privacy and regret: (i) a reformulation lemma mapping PrivITP's ReLU acceptance weights onto a randomized-threshold mechanism, enabling an n-independent ex-post DP bound (Lemma C.1); (ii) a smoothed-weight policy comparison lemma unifying regret analysis across threshold noise, query noise, and finite-sample effects (Lemma C.3); and (iii) a regret-optimality proof (Theorem 4.3) showing PrivITP matches Huang et al. (2025) skyline up to an $R_{\max}\sigma$ noise-inflation term that cleanly separates alignment error from privacy cost.

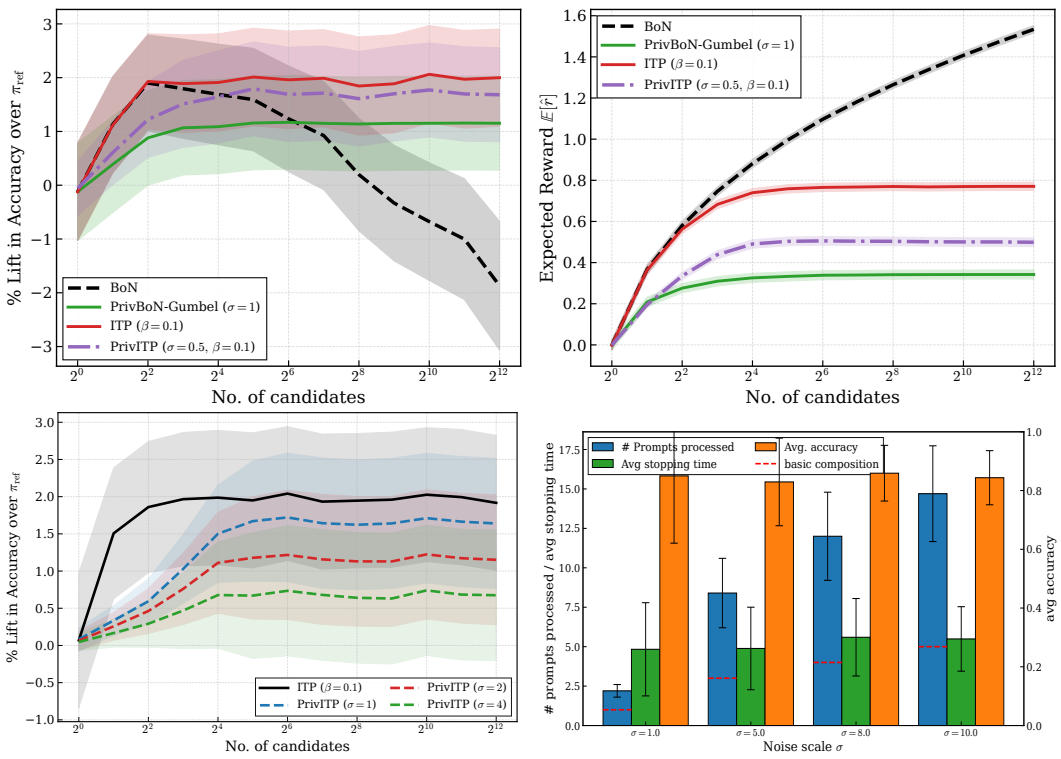


Figure 1: Experiments for Phi-3-Mini-Instruct policy and Gemma-RM on the GSM8K test split. *Top row* demonstrates reward hacking in BoN and its mitigation by PrivBoN, ITP, and PrivITP. *Bottom left* shows privacy vs. utility trade-off in ITP: low noise (low privacy) leads to better accuracy, and vice versa. *Bottom right* demonstrates the advantage of FSRC composition in PrivITP's ex-post privacy bound compared to the standard composition in ex-ante bound.

4.1 Multi-Query Deployment

In deployment, an aligned language model answers a stream of T prompts and produces a transcript of responses $(y^{*(1)}, \ldots, y^{*(T)})$. Each query consumes some privacy budget; the question is how to allocate this budget across queries to maximize alignment quality over the lifetime of the deployment. PrivITP's ex-post privacy guarantee (Theorem 4.1) provides a structural advantage: the privacy cost of each query depends on the actual halting time t, not on the worst-case budget n. When rejection sampling accepts early – the typical case for informative reward models, where $\mathbb{E}[t] = O(1)$ – the per-query privacy cost is much smaller than the worst-case bound. The FSRC framework of Lebensold et al. (2024, Alg. 2) is designed to exploit this structure: it composes mechanisms with ex-post privacy guarantees and halts when the cumulative privacy cost exceeds the total budget.

Proposition 4.5 (PrivITP composition via FSRC)**.** *Algorithm 3 satisfies $(\varepsilon_{\text{total}}, \delta)$-DP under fully-adaptive query selection. The expected number of queries answered before halting is $\mathbb{E}[T^*] \geq \frac{\varepsilon_{\text{total}}}{\mathbb{E}[\varepsilon_\tau]}$, where $\mathbb{E}[\varepsilon_\tau]$ is the expected per-query privacy cost under the distribution of halting times t_τ induced by the noisy reward distribution.*

In the favorable regime where $\mathbb{E}[\varepsilon_\tau] \ll \varepsilon_{\max}$, where $\epsilon_{\max} := \epsilon_1 + \epsilon_2^{\text{post}}(n)$ is the worst-case per-query cost, FSRC-composed PrivITP answers $\Theta(\varepsilon_{\max}/\mathbb{E}[\varepsilon_\tau])$ times more queries than standard composition for the same total budget. FSRC strictly dominates privacy filters (Rogers et al., 2023), which must charge the worst-case per-query cost rather than the realized ex-post cost (see Remark D.1)

5 Experiments

We validate our theoretical results following the experimental setup of Huang et al. (2025): four reward models (Oasst (1.4B), Gemma (2B), Llama (3B), and Armo (7B)) and two base LMs (Phi3-Mini-Instruct and Gemma-2-2B-Instruct) evaluated on test splits of GSM8K (Cobbe et al., 2021), MATH-500 (Hendrycks et al., 2021), and math, chemistry splits of MMLU (Hendrycks et al., 2020).

Table 1: Percentage lift in accuracy over base policy Phi-3-Mini-Instruct at $n = 2^{12}$ candidates. Values are mean $\pm$ standard error over all prompts across the GSM8K, MMLU, and MATH test splits. Same (RM, dataset) configuration uses the same noise σ for private variants and the same regularization β for ITP variants.

Dataset	Algorithm	Oasst-RM	Gemma-RM	Llama-RM	Armo-RM
GSM8K	BoN	-5.18 ± 1.23	-1.87 ± 1.20	7.64 ± 0.82	6.47 ± 0.86
	PrivBoN	0.32 ± 0.91	1.15 ± 0.87	6.90 ± 0.76	5.78 ± 0.78
	PrivITP	0.81 ± 0.89	1.68 ± 0.87	7.46 ± 0.78	6.63 ± 0.78
	ITP	0.85 ± 0.90	2.00 ± 0.90	7.61 ± 0.79	6.81 ± 0.79
MMLU	BoN	2.54 ± 2.74	5.45 ± 2.94	-12.74 ± 2.92	8.40 ± 4.37
	PrivBoN	2.32 ± 2.22	4.22 ± 2.25	1.88 ± 2.24	3.37 ± 3.22
	PrivITP	3.80 ± 2.28	6.09 ± 2.33	2.76 ± 2.29	5.59 ± 3.40
	ITP	5.09 ± 2.35	7.35 ± 2.46	4.69 ± 2.40	8.24 ± 3.74
MATH	BoN	-5.21 ± 1.60	2.10 ± 1.79	10.80 ± 1.98	6.10 ± 1.91
	PrivBoN	1.33 ± 1.55	2.94 ± 1.61	9.14 ± 1.73	4.51 ± 1.63
	PrivITP	1.72 ± 1.58	3.38 ± 1.62	10.28 ± 1.79	7.04 ± 1.70
	ITP	1.64 ± 1.59	3.45 ± 1.63	10.68 ± 1.81	7.80 ± 1.77

For each prompt, we sample 10000 responses from base policy at temperature 1 with zero-shot CoT prompting (Wei et al., 2022), then bootstrap $M = 50$ replicates of each sample to run each of the four algorithms: BoN, PrivBoN, ITP, and PrivITP. Unless otherwise mentioned, private algorithms use the same noise σ (for PrivITP, we set $\sigma_X = \sigma_Z = \sigma/2$), and ITP variants use the same regularization β on the same (RM, dataset) configuration. Ablation studies on the noise level σ (equivalently, on privacy budget ϵ) and regularization β are presented in Appendix F.

Scaling behavior across algorithms (Theorems 3.3, 4.3). Figure 1(top row) shows % percentage lift in accuracy over base policy Phi3-Mini and expected proxy reward under Gemma-RM for GSM8K as we vary n. We observe (i) BoN's accuracy rises then falls with n, exhibiting reward hacking, while its proxy reward rises monotonically—the signature of overfitting to $\hat{r}_{\mathcal{D}}$; (ii) PrivBoN and PrivITP accuracies are monotone in n, confirming hacking mitigation; (iii) PrivITP matches or exceeds PrivBoN at the same noise level. Figure 1(bottom left) shows that PrivITP approaches the non-private ITP skyline as noise (and hence, privacy) decreases. Figures for other RMs are in Appendix F.

Generalization across RMs and datasets. At $n = 2^{12}$, we report the percentage lift in accuracy over Phi3-Mini-Instruct base policy for each (algorithm, RM, dataset) configuration in Table 2. PrivITP dominates PrivBoN on every (RM, dataset) configuration, and on GSM8K it recovers between 84% and 98% of the non-private ITP skyline across all four reward models. On MMLU and MATH, PrivITP again closely tracks ITP within standard error. The reward-hacking signature in BoN is most pronounced for the weaker RMs (Oasst, Gemma), while PrivITP recovers near-skyline performance in each case. For stronger RMs (Llama, Armo) on GSM8K and MATH, BoN remains competitive at $n = 2^{12}$, which indicates that reward-hacking pressure is delayed when the base policy has lower accuracy. Results for Gemma2-2B-Instruct base model are in Appendix F.

Multi-query deployment via FSRC (Proposition 4.5). To validate the composition advantage of PrivITP's ex-post structure, we compare the number of GSM8K queries answered before budget exhaustion under (a) FSRC-composed PrivITP, (b) PrivITP basic composition (worst-case per-query cost $\varepsilon_{\max} = \varepsilon_1 + \varepsilon_2^{\text{post}}(n)$) sweeping over noise σ. We fix $n = 16$, $\beta = 0.05$ and total privacy budget $\varepsilon_{\text{total}} = 50$, Under Gemma-RM and across 10 random seeds, Figure 1(bottom right) shows that FSRC answers $\sim 3\times$ more queries, with empirical halting time $t_\tau \ll n$, confirming Remark 4.2. Similar behavior is observed for other RMs, and plots are in Appendix F.

Additional results – open-ended evaluation on AlpacaEval-2.0 dataset (Li et al., 2023), wall-clock overhead of our algorithms, and a privacy-budget sensitivity analysis showing favorable trend – are reported in Appendix G. **Open directions** include extending to sequential and multi-turn test-time alignment (Yu et al., 2025) and connecting inference-time privacy to training-time private alignment (Yu et al., 2021) for end-to-end pipeline guarantees.

AI USE STATEMENT

In preparing this work, we used generative AI tools solely as general-purpose writing and coding assistants. Specifically, AI tools were used to (i) polish exposition and improve the clarity of prose, and (ii) assist with routine coding tasks such as generating plotting scripts and refactoring experiment code. All theoretical contributions—including the problem formulation, theorem statements, proof strategies, and complete proofs—were developed and verified by the authors without AI assistance. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-polished text was checked against the authors' intended meaning. We take full responsibility for the final content of this work, including all text, claims, proofs, and experimental artifacts. This use is consistent with the ICLR 2027 AI Policy for Authors.

ETHICS STATEMENT

This work studies differential privacy for inference-time alignment, and its primary aim is to *strengthen* the privacy protections afforded to the human preference data used to train reward models. Our methods add formal (ϵ, δ)-DP guarantees to a widely deployed selection procedure (Best-of-N) that currently offers no such protection, which we view as a net positive for the privacy of individuals whose preferences are captured in alignment datasets. Our experiments use only publicly available benchmarks (GSM8K, MATH, MMLU, and AlpacaEval-2.0) and publicly released reward and language models; no new human-subjects data was collected, and no personally identifiable information was used. We note two limitations relevant to responsible deployment: (i) our privacy guarantees rely on a correctly specified sensitivity bound Δ_r, and misspecification would weaken the stated ϵ; and (ii) PrivITP's ex-post guarantee is weaker against timing side-channels than an ex-ante bound. Beyond the risks already inherent to language-model alignment, we do not foresee direct paths to harmful misuse. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure our results are reproducible. All theoretical claims are stated with explicit assumptions in Sections 3, 3.1, and 4, and complete proofs are provided in Appendices B, C, and D, with the Gaussian-mechanism variant of PrivBoN analyzed in Appendix E. Our full experimental setup—base policies, reward models, datasets, sampling temperature, number of candidates, bootstrap replication, and the noise (σ) and regularization (β) hyperparameters—is described in Section 5 and detailed in Appendix F, with additional evaluations (open-ended generation, wall-clock overhead, and privacy-budget sensitivity) in Appendix G. All datasets and models used are publicly available and cited in the main text, and our privacy accounting, including the exact ϵ computation and sensitivity assumptions, is given in Appendix G. Anonymized source code to reproduce all experiments, tables, and figures is included in the supplementary materials.

REFERENCES

Gholamali Aminian, Idan Shenfeld, Amir R Asadi, Ahmad Beirami, and Youssef Mroueh. Best-of-n through the smoothing lens: Kl divergence and regret analysis. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*.

Tim Baumgärtner, Yang Gao, Dana Alon, and Donald Metzler. Best-of-venom: Attacking rlhf by injecting poisoned preference data. *arXiv preprint arXiv:2404.05530*, 2024.

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 *International Conference on Machine Learning*, pp. 3580–3602. PMLR, 2025.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

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

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

David Durfee and Ryan M Rogers. Practical differentially private top-k selection with pay-what-you-get composition. *Advances in Neural Information Processing Systems*, 32, 2019.

Cynthia Dwork and Vitaly Feldman. Privacy-preserving prediction. In *Conference On Learning Theory*, pp. 1693–1702. PMLR, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Qizhang Feng, Siva Rajesh Kasa, Santhosh Kumar Kasa, Hyokun Yun, Choon Hui Teo, and Sravan Babu Bodapati. Exposing privacy gaps: Membership inference attack on preference data for llm alignment. *arXiv preprint arXiv:2407.06443*, 2024.

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

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In *International Conference on Machine Learning*, pp. 25075–25126. PMLR, 2025.

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

Seungone Kim, Jemin Shin, Yejin Cho, Joel Jang, Shayne Lu, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

Jonathan Lebensold, Doina Precup, and Borja Balle. On the privacy of selection mechanisms with gaussian noise. In *International Conference on Artificial Intelligence and Statistics*, pp. 1495–1503. PMLR, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The twelfth international conference on learning representations*, 2023.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2007.

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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Ryan M Rogers, Gennady Samorodnitsk, Steven Z Wu, and Aaditya Ramdas. Adaptive privacy composition for accuracy-first mechanisms. *Advances in Neural Information Processing Systems*, 36:15833–15854, 2023.

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.

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.

Florian Tramèr, Gautam Kamath, and Nicholas Carlini. Position: Considerations for differentially private learning with large-scale public pretraining. *arXiv preprint arXiv:2212.06470*, 2022.

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.

Jiongxiao Wang, Junlin Wu, Muhao Chen, Yevgeniy Vorobeychik, and Chaowei Xiao. Rlhfpoison: Reward poisoning attack for reinforcement learning with human feedback in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2551–2570, 2024.

Liangyu Wang, Junxiao Wang, Jie Ren, Zihang Xiang, David E Keyes, and Di Wang. Private training large-scale models with efficient dp-sgd. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Bosma Maarten, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24824–24837, 2022.

Justin Whitehouse, Aaditya Ramdas, Ryan Rogers, and Steven Wu. Fully-adaptive composition in differential privacy. In *International conference on machine learning*, pp. 36990–37007. PMLR, 2023.

Junlin Wu, Jiongxiao Wang, Chaowei Xiao, Chenguang Wang, Ning Zhang, and Yevgeniy Vorobeychik. Preference poisoning attacks on reward model learning. In *2025 IEEE Symposium on Security and Privacy (SP)*, pp. 1622–1640. IEEE, 2025.

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.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Yue Yu, Qiwei Di, Quanquan Gu, and Dongruo Zhou. On the limits of test-time compute: Sequential reward filtering for better inference. *arXiv preprint arXiv:2512.04558*, 2025.

Yizhou Zhang, Kishan Panaganti, Laixi Shi, Juba Ziani, and Adam Wierman. Kl-regularization itself is differentially private in bandits and rlhf. *arXiv preprint arXiv:2505.18407*, 2025.

Xingyu Zhou, Yulian Wu, and Francesco Orabona. A unified theoretical analysis of private and robust offline alignment: from rlhf to dpo. In *International Conference on Machine Learning*, pp. 79366–79388. PMLR, 2025a.

Xingyu Zhou, Yulian Wu, Wenqian Weng, and Francesco Orabona. Square χpo: Differentially private and robust χ^2-preference optimization in offline direct alignment. In *International Conference on Machine Learning*, pp. 79389–79413. PMLR, 2025b.

A Related work

Best-of-N alignment. Beirami et al. (2025) and Yang et al. (2024) analyze BoN's KL-reward tradeoff. Verdun et al. (2025) introduce Soft BoN (SBoN), showing $O(1/n)$ convergence to the KL-regularized optimum. Aminian et al. derive KL and regret bounds for SBoN under proxy reward error. Our MBON mechanisms are algorithmically equivalent to SBoN; the novelty lies in the privacy derivation of σ, the regret optimality proof, and the PrivITP extension. Khalaf et al. characterize inference-time reward hacking; our reward hacking bounds (Section 3.2) complement their results by quantifying how noise injection caps the hacking gap. Yu et al. (2025) show that sequential test-time compute dominates parallel methods such as BoN; extending PrivITP to the sequential setting is an open direction. Huang et al. (2025) establish the regret skyline $\Omega(\sqrt{C^{\pi^*}\varepsilon_{\mathrm{RM}}^2})$ and introduce InferenceTimePessimism, a χ^2-regularized rejection sampling algorithm that achieves this bound.

Differential privacy in LLM alignment. All prior work on private alignment operates at *training time*. Zhang et al. (2025) prove that KL-regularization itself provides DP for the trained policy. Zhou et al. (2025a) study private and robust offline RLHF under label DP. Wang et al. (2025) propose DP-AdamW for private DPO fine-tuning. Zhou et al. (2025b) gives private and robust algorithms for training-based alignment. Prior work on private inference has studied differentially private prediction (Dwork & Feldman, 2018), private top-k selection (Durfee & Rogers, 2019), but the specific interaction between privacy noise and alignment quality in inference-time methods has not been analyzed. Our work is the first to provide formal privacy guarantees at *inference time*, protecting the reward model's training data during deployment.

B Proofs for Private Best-of-N

We provide detailed proofs for every result stated in the main text.

B.1 Proof of Theorem 3.1: PrivBoN Privacy

Proof. Condition on the candidate set $y_{1:n}$, drawn i.i.d. from $\pi_0(\cdot|x)$ independently of D. The Gumbel-max trick states that if $g_1, \ldots, g_n \overset{\text{iid}}{\sim} \mathrm{Gumbel}(0, \sigma)$, then:

$$\Pr\left(\arg\max_i(\hat{r}_D(x, y_i) + g_i) = k\right) = \frac{e^{\hat{r}_D(x,y_k)/\sigma}}{\sum_{j=1}^n e^{\hat{r}_D(x,y_j)/\sigma}}.$$

This is precisely the exponential mechanism with score function $q(k; D) = \hat{r}_D(x, y_k)$ and privacy parameter $\varepsilon/(2\Delta_r) = 1/\sigma$.

For adjacent $D \sim D'$, any candidate y_k, and the fixed set $y_{1:n}$:

$$\frac{\Pr(i^* = k \mid D)}{\Pr(i^* = k \mid D')} = \frac{e^{\hat{r}_D(x,y_k)/\sigma}}{e^{\hat{r}_{D'}(x,y_k)/\sigma}} \cdot \frac{\sum_j e^{\hat{r}_{D'}(x,y_j)/\sigma}}{\sum_j e^{\hat{r}_D(x,y_j)/\sigma}}. \tag{7}$$

The first factor is at most $e^{\Delta_r/\sigma} = e^{\varepsilon/2}$ (since $|\hat{r}_D - \hat{r}_{D'}| \le \Delta_r$). For the second factor, each term in the numerator sum is at most $e^{\Delta_r/\sigma}$ times the corresponding denominator term, so the ratio is at most $e^{\Delta_r/\sigma} = e^{\varepsilon/2}$. Combined: the likelihood ratio is at most e^{ε}.

Since the output y_{i^*} is a deterministic function of i^* and $y_{1:n}$ (post-processing), and $y_{1:n}$ is independent of D, the overall mechanism is ε-DP. □

B.2 Proof of Proposition 3.2: PrivBoN Closed-Form Policy

Proof. For brevity, we denote $r(x, y) := \hat{r}_D(x, y)$. Note that the event $\{y^* = y\}$ decomposes as $\bigcup_{i=1}^n \{i^* = i, y_i = y\}$. These are disjoint (only one index wins), so:

$$\pi_{\mathrm{PrivBoN}}(y|x) = \mathbb{E}_{y_{1:n}}\left[\sum_{i=1}^n \Pr(i^* = i \mid y_{1:n})\, \mathbf{1}\{y_i = y\}\right].$$

By the i.i.d. structure of $y_1, \ldots, y_n$ and symmetry, all n terms are identically distributed. Hence:

$$= n\, \mathbb{E}_{y_{1:n}}\left[\Pr(i^* = 1 \mid y_{1:n})\, \mathbf{1}\{y_1 = y\}\right].$$

Apply the tower property, conditioning on y_1:

$$= n\,\mathbb{E}_{y_1}\left[\mathbf{1}\{y_1 = y\}\,\mathbb{E}_{y_{2:n}}\left[\frac{e^{r(x,y_1)/\sigma}}{\sum_{j=1}^n e^{r(x,y_j)/\sigma}}\right]\right],$$

where inside the inner expectation, we have used the probability of selecting candidate 1 using the Gumbel-max trick.

Now, since $y_1 \sim \pi_0(\cdot|x)$, we have $\mathbb{E}_{y_1}[\mathbf{1}\{y_1 = y\}f(y_1)] = \pi_0(y|x)\,f(y)$. Applying this identity yields the stated formula.

BoN Recovery ($\sigma \to 0$) Note that $\pi_{\mathrm{PrivBoN}}(y|x) = n\,\pi_0(y|x)\,\mathbb{E}_{y_{2:n}}\left[\frac{e^{r_1/\sigma}}{e^{r_1/\sigma}+\sum_{j=2}^n e^{r_j/\sigma}}\right]$, where $r_1 = r(x,y)$ and $r_j = r(x,y_j)$ for $j \geq 2$. Inside the expectation, consider the softmax share for candidate 1:

$$\psi(r_1) = \frac{e^{r_1/\sigma}}{e^{r_1/\sigma} + \sum_{j=2}^n e^{r_j/\sigma}} = \frac{1}{1 + \sum_{j=2}^n e^{(r_j - r_1)/\sigma}}.$$

Let $M = \max_{j\geq 2} r_j$. Factor out $e^{(M-r_1)/\sigma}$ to get

$$\psi(r_1) = \frac{1}{1 + e^{(M-r_1)/\sigma}\sum_{j=2}^n e^{(r_j - M)/\sigma}}.$$

If $r_1 > M$ (i.e., r_1 is the unique maximum), then $(M - r_1)/\sigma \to -\infty$ when $\sigma \to 0$, so the softmax share $\psi(r_1) =\to 1$. If $r_1 < M$, then $(M - r_1)/\sigma \to +\infty$, so the softmax share $\psi(r_1) =\to 0$. If $r_1 = M$ (tie), then the limit depends on the number of ties, but ties have probability zero under continuous reward distributions.

Therefore, the softmax share satisfies

$$\lim_{\sigma\to 0}\psi(r_1) = \lim_{\sigma\to 0}\frac{e^{r_1/\sigma}}{\sum_{j=1}^n e^{r_j/\sigma}} = \mathbf{1}\{r_1 \geq \max_{j\geq 2} r_j\}$$

Taking expectation, we get

$$\mathbb{E}_{y_{2:n}}[\mathbf{1}\{r(x,y) \geq \max_{j\geq 2} r(x,y_j)\}] = \prod_{j=2}^n \Pr(r(x,y) \geq r(x,y_j)) = F_{\pi_0}(y|x)^{n-1},$$

where $F_{\pi_0}(\cdot|x)$ is the cumulative distribution function of $r(x,Y)$ where $Y \sim \pi_0(\cdot|x)$. Hence $\lim_{\sigma\to 0}\pi_{\mathrm{PrivBoN}}(y|x) = n\,\pi_0(y|x)\,F_{\pi_0}(y|x)^{n-1} = \pi_{\mathrm{BoN}}(y|x)$.

PrivBoN Policy Limit ($n \to \infty$) Again, consider the softmax share for candidate 1 inside the expectation:

$$\psi(r_1) = \frac{e^{r_1/\sigma}}{e^{r_1/\sigma} + \sum_{j=2}^n e^{r_j/\sigma}} = \frac{1}{1 + \sum_{j=2}^n e^{(r_j - r_1)/\sigma}}.$$

Define $S_n = \frac{1}{n-1}\sum_{j=2}^n e^{r_j/\sigma}$. By the strong law of large numbers, $S_n \xrightarrow{\text{a.s.}} M(\sigma) = \mathbb{E}_{Y\sim\pi_0(\cdot|x)}[e^{r(x,Y)/\sigma}]$. Hence

$$\psi(r_1) \to \frac{1}{1 + (n-1)M(\sigma)e^{-r_1/\sigma}} \implies n\,\psi(r_1) \to n\cdot e^{r_1/\sigma}/((n-1)M(\sigma) + e^{r_1/\sigma}).$$

Hence $n\,\psi(r_1) \to e^{r_1/\sigma}/M(\sigma)$ as $n \to \infty$. Therefore $\pi_{\mathrm{PrivBoN}}(y|x) \to \pi_0(y|x)\cdot e^{r(x,y)/\sigma}/M(\sigma)$. □

B.3 Proof of Theorem 3.3: PrivBoN Regret

Proof. Define $r(x,y) := \hat{r}_D(x,y)$. We decompose the regret via the limiting tilted policy $\pi_\infty(y|x) = \pi_0(y|x)e^{r(x,y)/\sigma}/M(\sigma)$, where $M(\sigma) = \mathbb{E}_{\pi_0}[e^{r(x,y)/\sigma}]$:

$$\mathrm{Reg}(\pi_{\mathrm{PrivBoN}}; x) = \underbrace{J(\pi^*; x) - J(\pi_\infty; x)}_{\text{(A)}} + \underbrace{J(\pi_\infty; x) - J(\pi_{\mathrm{PrivBoN}}; x)}_{\text{(B)}}. \tag{8}$$

Step 1: Bounding (A). Regret of PrivBoN under infinite samples.

Define the shorthand $\mathbb{E}_\pi[f] := \mathbb{E}_{y\sim\pi(\cdot|x)}[f(x,y)]$ for any function $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ and $\mathrm{KL}(\pi\|\pi') := \mathrm{KL}(\pi(\cdot|x)\|\pi'(\cdot|x))$. Now, write the true reward in terms of the proxy: $r^*(x,y) = r(x,y) - \epsilon(x,y)$. Then

$$\begin{aligned}(\mathrm{A}) &= \mathbb{E}_{\pi^*}[r^*] - \mathbb{E}_{\pi_\infty}[r^*] \\ &= \big(\mathbb{E}_{\pi^*}[r] - \mathbb{E}_{\pi_\infty}[r]\big) - \big(\mathbb{E}_{\pi^*}[\epsilon] - \mathbb{E}_{\pi_\infty}[\epsilon]\big). \end{aligned} \tag{9}$$

Bounding the reward gap. The tilted policy π_∞ solves $\max_\pi\{\mathbb{E}_\pi[r] - \sigma\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\}$. By the optimality condition, for any policy π,

$$\mathbb{E}_\pi[r] - \mathbb{E}_{\pi_\infty}[r] \le \sigma\big(\mathrm{KL}(\pi\|\pi_0) - \mathrm{KL}(\pi_\infty\|\pi_0)\big) \le \sigma\,\mathrm{KL}(\pi\|\pi_0).$$

Applying this with $\pi = \pi^*$, we get

$$\mathbb{E}_{\pi^*}[r] - \mathbb{E}_{\pi_\infty}[r] \le \sigma\,\mathrm{KL}(\pi^*\|\pi_{\mathrm{ref}}).$$

Since $\mathrm{KL}(\pi^*\|\pi_{\mathrm{ref}}) \le \chi^2(\pi^*\|\pi_{\mathrm{ref}}) = \frac{1}{2}(C^{\pi^*}(x) - 1)$ (by the standard inequality $\mathrm{KL} \le \chi^2$), we obtain

$$\mathbb{E}_{\pi^*}[r] - \mathbb{E}_{\pi_\infty}[r] \le \frac{\sigma}{2}(C^{\pi^*}(x) - 1). \tag{10}$$

Bounding the error terms. For any policy π with coverage $C^\pi(x)$, the Cauchy–Schwarz inequality gives:

$$\begin{aligned}|\mathbb{E}_\pi[\epsilon]| = \left|\mathbb{E}_{\pi_0}\left[\frac{\pi(y|x)}{\pi_0(y|x)}\epsilon(x,y)\right]\right| &\le \sqrt{\mathbb{E}_{\pi_0}\left[\left(\frac{\pi(x,y)}{\pi_0(x,y)}\right)^2\right]\mathbb{E}_{\pi_0}[\epsilon^2(x,y)]} \\ &= \sqrt{\mathbb{E}_\pi\left[\frac{\pi(x,y)}{\pi_0(x,y)}\right]\mathbb{E}_{\pi_0}[\epsilon^2(x,y)]} = \sqrt{C^\pi(x)}\,\varepsilon_{\mathrm{RM}}(x).\end{aligned}$$

Applying this to π^* directly, we get

$$|\mathbb{E}_{\pi^*}[\epsilon]| \le \sqrt{C^{\pi^*}(x)}\,\varepsilon_{\mathrm{RM}}(x). \tag{11}$$

Now $\epsilon(x,y) \sim \mathcal{N}(0,\sigma_r^2(x))$ under π_0. So, the tilted distribution of $\epsilon(x,y)$ under π_∞ has density $\mathcal{N}(\sigma_r^2(x)/\sigma, \sigma_r^2(x))$. Therefore we get

$$\mathbb{E}_{y\sim\pi_\infty(\cdot|x)}[\epsilon(x,y)] = \frac{\sigma_r^2(x)}{\sigma}. \tag{12}$$

Combining (9), (10), (11), and (12):

$$(\mathrm{A}) \le \frac{\sigma}{2}(C^{\pi^*} - 1) + \sqrt{C^{\pi^*}}\,\varepsilon_{\mathrm{RM}} + \frac{\varepsilon_{\mathrm{RM}}^2}{\sigma}\,. \tag{13}$$

Bounding (B): Finite-sample gap. Let $w_k = e^{r(x,y_k)/\sigma}$ and $f_k = r^*(x,y_k)$ for candidates $y_1,\dots,y_n \sim \pi_0$. Assume that $r(x,y) \in [a,b]$ with range $c = b - a$, we have $w_k \in [e^{a/\sigma}, e^{b/\sigma}]$. Since $r^* \in [0, R_{\max}]$, we have $f_k \in [0, R_{\max}]$. By the Gumbel-max trick, the selected candidate y_{i^*} is drawn from the softmax distribution $\Pr[i^* = k \mid y_{1:n}] = w_k / \sum_j w_j$, so

$$J(\hat\pi_{\mathrm{PrivBoN}}; x) = \mathbb{E}_{\pi_{\mathrm{PrivBoN}}}[r^*] = \mathbb{E}_{y_{1:n}}\left[\frac{\sum_k w_k f_k}{\sum_k w_k}\right] = \mathbb{E}\left[\frac{\hat Z_n^{(r)}}{\hat Z_n}\right],$$

where $\hat Z_n = \frac{1}{n}\sum_k w_k$ and $\hat Z_n^{(r)} = \frac{1}{n}\sum_k w_k f_k$ are empirical means, with population counterparts $Z = \mathbb{E}_{\pi_0}[w] \ge e^{a/\sigma}$ and $Z^{(r)} = \mathbb{E}_{\pi_0}[wf]$, where $w = e^{r(x,y)/\sigma}$ and $f = r^*(x,y)$, so that $J(\pi_\infty; x) = Z^{(r)}/Z$.

Let $A = \hat Z_n^{(r)} - Z^{(r)}$ and $B = \hat Z_n - Z$. Then:

$$\frac{\hat Z_n^{(r)}}{\hat Z_n} - \frac{Z^{(r)}}{Z} = \frac{\hat Z_n^{(r)} Z - Z^{(r)}\hat Z_n}{Z\hat Z_n} = \frac{AZ - Z^{(r)}B}{Z\hat Z_n}.$$

By the triangle inequality and $Z^{(r)} = \mathbb{E}_{\pi_0}[wr^*] \le R_{\max}\mathbb{E}_{\pi_0}[w] = R_{\max}Z$ (since $r^* \in [0, R_{\max}]$ and $w > 0$), we have

$$|AZ - Z^{(r)}B| \le Z|A| + Z^{(r)}|B| \le Z(|A| + R_{\max}|B|),$$

giving $|\hat{Z}_n^{(r)}/\hat{Z}_n - Z^{(r)}/Z| \le (|A| + R_{\max}|B|)/\hat{Z}_n$.

Define $\mathcal{E} = \{\hat{Z}_n \ge Z/2\}$. Since $w_k \in [e^{a/\sigma}, e^{b/\sigma}]$ are i.i.d. with mean Z, Hoeffding's inequality gives

$$\Pr[\mathcal{E}^c] = \Pr[\hat{Z}_n - Z < -Z/2] \le \exp\left(-\frac{2n(Z/2)^2}{(e^{b/\sigma} - e^{a/\sigma})^2}\right) \le \exp\left(-\frac{nZ^2}{2e^{2b/\sigma}}\right),$$

where the last inequality uses $e^{b/\sigma} - e^{a/\sigma} \le e^{b/\sigma}$.

Bound on the good event. On $\mathcal{E}$, $\hat{Z}_n \ge Z/2$, so $(|A| + R_{\max}|B|)/\hat{Z}_n \le 2(|A| + R_{\max}|B|)/Z$. Now, via Jensen inequality ($\mathbb{E}|X| \le \sqrt{\mathbb{E}[X^2]}$), we get

$$\mathbb{E}[|B|] \le \sqrt{\mathbb{E}[(\hat{Z}_n - Z)^2]} = \sqrt{\mathrm{Var}(\hat{Z}_n)} = \frac{\sqrt{\mathrm{Var}_{\pi_0}(w)}}{\sqrt{n}} \le \frac{\sqrt{\mathbb{E}_{\pi_0}[w^2]}}{\sqrt{n}} \le \frac{e^{b/\sigma}}{\sqrt{n}},$$

using $\mathrm{Var}(w) \le \mathbb{E}[w^2]$ and $w \le e^{b/\sigma}$. Similarly

$$\mathbb{E}[|A|] \le \sqrt{\mathrm{Var}(\hat{Z}_n^{(r)})} = \frac{\sqrt{\mathrm{Var}_{\pi_0}(wr^*)}}{\sqrt{n}} \le \frac{\sqrt{\mathbb{E}_{\pi_0}[w^2(r^*)^2]}}{\sqrt{n}} \le \frac{R_{\max}e^{b/\sigma}}{\sqrt{n}},$$

using $(r^*)^2 \le R_{\max}^2$. Therefore, we get

$$\mathbb{E}\left[\left|\frac{\hat{Z}_n^{(r)}}{\hat{Z}_n} - \frac{Z^{(r)}}{Z}\right| \cdot \mathbf{1}_{\mathcal{E}}\right] \le \frac{2}{Z}\left(\frac{R_{\max}e^{b/\sigma}}{\sqrt{n}} + \frac{R_{\max}e^{b/\sigma}}{\sqrt{n}}\right) = \frac{4R_{\max}e^{b/\sigma}}{Z\sqrt{n}}.$$

Bound on the bad event. Both $\hat{Z}_n^{(r)}/\hat{Z}_n$ and $Z^{(r)}/Z$ are weighted averages of values in $[0, R_{\max}]$ (the former is a convex combination of $f_k \in [0, R_{\max}]$; the latter is $\mathbb{E}[wf]/\mathbb{E}[w]$ with $f \in [0, R_{\max}]$). Hence, their difference is bounded by $R_{\max}$. Thus

$$\mathbb{E}\left[\left|\frac{\hat{Z}_n^{(r)}}{\hat{Z}_n} - \frac{Z^{(r)}}{Z}\right| \cdot \mathbf{1}_{\mathcal{E}^c}\right] \le R_{\max} \cdot \Pr[\mathcal{E}^c] \le R_{\max}\exp\left(-\frac{nZ^2}{2e^{2b/\sigma}}\right).$$

By the triangle inequality for expectations, we obtain

$$|J(\pi_\infty; x) - J(\pi_{\mathrm{PrivBoN}}; x)| \le \mathbb{E}\left[\left|\frac{\hat{Z}_n^{(r)}}{\hat{Z}_n} - \frac{Z^{(r)}}{Z}\right|\right] \le \frac{4R_{\max}e^{b/\sigma}}{Z\sqrt{n}} + R_{\max}\exp\left(-\frac{nZ^2}{2e^{2b/\sigma}}\right),$$

where $Z = \mathbb{E}_{\pi_0}[e^{r/\sigma}] \ge e^{a/\sigma}$. Using $Z \ge e^{a/\sigma}$, this simplifies to:

$$|J(\pi_\infty) - J(\hat{\pi}_{\mathrm{MBON}})| \le \frac{4R_{\max}e^{c/\sigma}}{\sqrt{n}} + R_{\max}\exp\left(-\frac{n}{2e^{2c/\sigma}}\right) \le \frac{4R_{\max}e^{c/\sigma}}{\sqrt{n}} + \varepsilon_{\mathrm{RM}} \tag{14}$$

for $n \ge 2e^{2c/\sigma}\log(R_{\max}/\varepsilon_{\mathrm{RM}})$, where $c = b - a$ is the reward range.

Adding (13) and (14), we get

$$\mathrm{Reg}^{\pi^*}(x) \le \sigma(C^{\pi^*} - 1) + \frac{\varepsilon_{\mathrm{RM}}^2}{\sigma} + \sqrt{C^{\pi^*}}\,\varepsilon_{\mathrm{RM}} + O\left(\frac{R_{\max}e^{c/\sigma}}{\sqrt{n}}\right).$$

The regret bound is modulo the assumption that $r(x, y) \in [a, b]$ with range $c = b - a$. Under the Gaussian error model $r(x, y) = r^*(x, y) + \epsilon(x, y)$ with $\epsilon(x, y) \sim \mathcal{N}(0, \sigma_r^2(x))$, the reward is unbounded. We handle this by conditioning on the high-probability event $\mathcal{F} := \{\max_k |\epsilon(x, y_k)| \le \sigma_r\sqrt{2\log(n/\Delta_r)}\}$, which has $\Pr[\mathcal{F}] \ge 1 - \Delta_r$ by Gaussian tail and union bound. On $\mathcal{F}$, the rewards lie in $[-\sigma_r\sqrt{2\log(n/\Delta_r)},\ R_{\max} + \sigma_r\sqrt{2\log(n/\Delta_r)}]$, so the bounded analysis applies with

effective range $c = R_{\max} + 2\sigma_r\sqrt{2\log(n/\Delta_r)}$. On $\mathcal{F}^c$, the regret is bounded trivially by $R_{\max}$, contributing $R_{\max}\Delta_r$. Setting $\Delta_r = 1/\sqrt{n}$, the effective range becomes $c = O(R_{\max} + \sigma_r\sqrt{\log n})$. Substituting this, we get

$$\mathrm{Reg}^{\pi^*}(x) \le \sigma(C^{\pi^*} - 1) + \frac{\varepsilon_{\mathrm{RM}}^2}{\sigma} + \sqrt{C^{\pi^*}}\,\varepsilon_{\mathrm{RM}} + O\left(\frac{R_{\max}e^{(R_{\max}+\varepsilon_{\mathrm{RM}}\sqrt{\log n})/\sigma}}{\sqrt{n}}\right)$$

for $n = \Omega\left(e^{(R_{\max}+\varepsilon_{\mathrm{RM}}\sqrt{\log n})/\sigma}\log(R_{\max}/\varepsilon_{\mathrm{RM}}\right)$. □

B.4 On the Gaussian-error assumption in Theorem 3.3

The Gaussian-error assumption in Theorem 3.3 gives the closed form $\mathbb{E}_{\pi_\infty}[\epsilon(x,y)] = \sigma_r^2(x)/\sigma$, which we use to derive the exact $\epsilon_{\mathrm{RM}}^2/\sigma$ overoptimization term. Under a general mean-squared-error assumption $\mathbb{E}_{\pi_0}[\epsilon(x,y)^2] = \epsilon_{\mathrm{RM}}^2(x)$, an application Cauchy–Schwarz inequality yields only $|\mathbb{E}_{\pi_\infty}[\epsilon]| \le \sqrt{C^{\pi_\infty}(x)}\,\epsilon_{\mathrm{RM}}(x)$, where C^{π_∞} is the coverage of the tilted policy. Since C^{π_∞} scales as $\exp(R_{\max}/\sigma)$ in the worst case, the overoptimization term inflates, weakening the "privacy is free" threshold.

The finite-sample analysis conditions on a high-probability boundedness event via a Gaussian tail bound (the union bound $\Pr[\max_k |\epsilon(x,y_k)| \le \sigma_r\sqrt{2\log(n/\delta)}] \ge 1-\delta$). For sub-Gaussian errors with parameter σ_r the analysis goes through unchanged up to constants. For bounded rewards $|\hat{r}_D| \le B$, the tail step becomes trivial, but B replaces $\sigma_r\sqrt{\log n}$ in the effective reward range, tightening the finite-sample term when $B \ll \sigma_r\sqrt{\log n}$ and loosening it otherwise.

The qualitative conclusions—bias-variance trade-off in σ, existence of an ϵ^* threshold above which privacy is free, and regret going down with n—hold under the general mean-squared-error assumption. The clean form of the threshold $\epsilon^*(x) = 2\Delta_r\sqrt{C^{\pi^*}(x)}/\epsilon_{\mathrm{RM}}(x)$, however, is specific to the Gaussian case.

B.5 Proof of Proposition 3.7 (Win-Rate Bound and Reward hacking)

Proof. **(a) Win-rate.** For brevity, we denote $r(x,y) := \hat{r}_D(x,y)$. Define $r_i = r(x,y_i)$ for $i \in [n]$ and $r_0 = r(x,y_0)$. Let $p_i = \Pr(r_i \ge r_0 \mid y_i)$ where $y_0, y_i \sim \pi_0(\cdot|x)$ independently.

We know that as $n \to \infty$, $\pi_{\mathrm{PrivBoN}} \to \pi_\infty$ where $\pi_\infty(y|x) \propto \pi_0(y|x)e^{r(x,y)/\sigma}$.

For Gaussian rewards $r(x,y) \sim \mathcal{N}(r^*(x,y), \sigma_r^2(x))$ under π_0. Now, the tilted distribution of $r(x,y)$ under π_∞ has density $\mathcal{N}(r^*(x,y) + \sigma_r^2(x)/\sigma, \sigma_r^2(x))$. Hence $Y \sim \pi_\infty(\cdot|x)$ has reward $r_Y \sim \mathcal{N}(r^*(x,y) + \sigma_r^2(x)/\sigma, \sigma_r^2(x))$ and $Z \sim \pi_0(\cdot|x)$ has reward $r_Z \sim \mathcal{N}(r^*(x,y), \sigma_r^2(x))$. Then:

$$r_Y - r_Z \sim \mathcal{N}(\sigma_r^2(x)/\sigma, 2\sigma_r^2(x)).$$

So $\mathrm{WinRate}_\infty^{\mathrm{PrivBoN}} = \Pr(r_Y \ge r_Z) = \Phi(\sigma_r^2(x)/(\sigma\sqrt{2}\sigma_r(x))) = \Phi(\sigma_r(x)/(\sigma\sqrt{2})) < 1$.

(b) Reward hacking. Denote $r(x,y) := \hat{r}_D(x,y)$, $r_i := r(x,y_i)$ and $\epsilon_i := \epsilon(x,y_i)$. For any fixed candidate set $y_{1:n}$, BoN selects $i^* = \arg\max_i r_i$. This maximizes $\mathbb{E}[r_{i^*}]$ over all (possibly randomized) selection rules, since the deterministic $\arg\max$ is optimal. PrivBoN's softmax selection is a specific randomized rule, hence suboptimal: $\mathbb{E}_{y\sim\pi_{\mathrm{PrivBoN}}(\cdot|x)}[r(x,y)] \le \mathbb{E}_{\pi_{\mathrm{BoN}}(\cdot|x)}[r(x,y)]$.

For BoN, the hacking gap is $\mathbb{E}_{y\sim\pi_{\mathrm{BoN}}(\cdot|x)}[\epsilon(x,y)] = \mathbb{E}_{y_{1:n}}[\sum_{i=1}^n \mathbf{1}\{i^* = i\}\epsilon_i] \le \mathbb{E}_{y_{1:n}}[\max_{i=1}^n \epsilon_i]$. Since each $\epsilon_i \sim \mathcal{N}(0, \sigma_r^2(x))$ under π_0 and independent, we have $\mathbb{E}[\max_i \epsilon_i] = O(\sigma_r(x)\sqrt{\log n})$ by standard results on Gaussian maximum.

In the limit $n \to \infty$, we have $\pi_{\mathrm{PrivBoN}} \to \pi_\infty$, where $\pi_\infty(y|x) \propto \pi_0(y|x)e^{r(x,y)/\sigma}$. Now $\epsilon(x,y) \sim \mathcal{N}(0, \sigma_r^2(x))$ under π_0. So, the tilted distribution of $\epsilon(x,y)$ under π_∞ has density $\mathcal{N}(\sigma_r^2(x)/\sigma, \sigma_r^2(x))$. Therefore, the hacking gap of PrivBoN becomes $\mathbb{E}_{y\sim\pi_\infty(\cdot|x)}[\epsilon(x,y)] = \frac{\sigma_r^2(x)}{\sigma}$ in the limit $n \to \infty$. □

C Proofs for Private Inference-Time Pessimism

C.1 Proof of Theorem 4.1 (PrivITP Privacy)

Proof. We analyze the two phases separately and combine them via basic composition.

Phase 1. Define $\lambda^*(D)$ as the unique solution to

$$F_D(\lambda) := \frac{1}{n}\sum_{i=1}^{n} \mathrm{relu}\big(\beta^{-1}(\hat{r}_D(x, y_i) - \lambda)\big) = 1,$$

where the candidates $y_1, \ldots, y_n \sim \pi_0(\cdot|x)$ are drawn independently of D. The function F_D is continuous and strictly decreasing in λ, with $F_D(\lambda) \to \infty$ as $\lambda \to -\infty$ and $F_D(\lambda) = 0$ for $\lambda \geq \max_i \hat{r}_D(x, y_i)$, so the solution exists and is unique.

Sensitivity of λ^*. For adjacent datasets $D \sim D'$ and each candidate y_i, we have $|\hat{r}_D(x, y_i) - \hat{r}_{D'}(x, y_i)| \leq \Delta_r$. Since $\mathrm{relu}(\beta^{-1}(z - \lambda))$ is monotone increasing in z, this gives pointwise

$$\mathrm{relu}\big(\beta^{-1}(\hat{r}_D(x, y_i) - (\lambda + \Delta_r))\big) \leq \mathrm{relu}\big(\beta^{-1}(\hat{r}_{D'}(x, y_i) - \lambda)\big) \leq \mathrm{relu}\big(\beta^{-1}(\hat{r}_D(x, y_i) - (\lambda - \Delta_r))\big).$$

Averaging over i, we get

$$F_D(\lambda + \Delta_r) \leq F_{D'}(\lambda) \leq F_D(\lambda - \Delta_r).$$

Setting $\lambda = \lambda^*(D')$, so $F_{D'}(\lambda^*(D')) = 1 = F_D(\lambda^*(D))$, we obtain

- Upper bound: $1 \leq F_D(\lambda^*(D') - \Delta_r)$, so by monotonicity of F_D, $\lambda^*(D') - \Delta_r \leq \lambda^*(D)$, i.e., $\lambda^*(D') \leq \lambda^*(D) + \Delta_r$.
- Lower bound: $1 \geq F_D(\lambda^*(D') + \Delta_r)$, so $\lambda^*(D') + \Delta_r \geq \lambda^*(D)$, i.e., $\lambda^*(D') \geq \lambda^*(D) - \Delta_r$.

Combining both, we get $|\lambda^*(D) - \lambda^*(D')| \leq \Delta_r$.

Gaussian mechanism. Since $\lambda^* : \mathcal{D} \to \mathbb{R}$ is a scalar deterministic function with global sensitivity at most Δ_r, releasing $\tilde{\lambda} = \lambda^*(D) + \zeta$ with $\zeta \sim \mathcal{N}(0, \sigma_X^2)$ is the scalar Gaussian mechanism (Dwork & Roth, 2014), yielding (ε_1, δ)-DP with $\varepsilon_1 = \Delta_r\sqrt{2\ln(1.25/\delta)}/\sigma_X$.

Phase 2. Fix λ (released privately from Phase 1, treated as public for the ex-post analysis) and fresh candidates $y'_1, \ldots, y'_n \sim \pi_0(\cdot|x)$ (independent of D). Let $q_i := \hat{r}_D(x, y'_i) \in [0, R_{\max}]$. At each round, Phase 2 computes $\tilde{r}_i = q_i + g'_i$ with $g_i \sim \mathcal{N}(0, \sigma_Z^2)$, weight $w_i = \mathrm{relu}(\beta^{-1}(\tilde{r}_i - \tilde{\lambda}))$, and acceptance $\xi_i \sim \mathrm{Bernoulli}(\min(w_i/M, 1))$. The halting time is $t^* = \min\{i : \xi_i = 1\}$.

Reformulation as randomized threshold. We show rejection sampling is equivalent in distribution to a threshold check with a randomized effective threshold. Let $u_i \sim \mathrm{Unif}[0, 1]$ be fresh independent uniform noise.

Lemma C.1 (Reformulation)**.** $\Pr[\xi_i = 1 \mid \tilde{r}_i] = \Pr\big[\tilde{r}_i \geq \tilde{\lambda} + \beta M u_i \mid \tilde{r}_i\big]$.

Indeed: (a) if $\tilde{r}_i < \tilde{\lambda}$, both sides are zero; (b) if $\tilde{\lambda} \leq \tilde{r}_i \leq \tilde{\lambda} + \beta M$, the LHS equals $w_i/M = (\tilde{r}_i - \lambda)/(\beta M)$, and the RHS equals $\Pr[u_i \leq (\tilde{r}_i - \tilde{\lambda})/(\beta M)] = (\tilde{r}_i - \lambda)/(\beta M)$; (c) if $\tilde{r}_i > \tilde{\lambda} + \beta M$, both sides are one.

With this reformulation, Phase 2 accepts candidate i iff $g_i \geq \tilde{\lambda} + \beta M u_i - q_i$, where (g_i, u_i) are independent fresh noise, independent across rounds.

Halting probability. Since the noise is independent across rounds, the probability of halting at time t factorizes:

$$\Pr[t^* = t \mid D] = \prod_{i=1}^{t-1}(1 - p_i(D)) \cdot p_t(D),$$

where the per-round acceptance probability (marginalizing over both g_i and u_i) is

$$p_i(D) = \Pr[g_i \geq \tilde{\lambda} + \beta M u_i - q_i] = \mathbb{E}_{u_i}\left[\Phi\left(\frac{q_i - \tilde{\lambda} - \beta M u_i}{\sigma_Z}\right)\right],$$

using $\Pr[g_i \geq z] = \Phi(-z/\sigma_Z)$. Correspondingly,

$$1 - p_i(D) = \mathbb{E}_{u_i}\left[\Phi\left(\frac{\tilde{\lambda} + \beta M u_i - q_i}{\sigma_Z}\right)\right].$$

Ratio under coupling. Under the coupling where all noise variables (g_i, u_i) are shared between D and D', let $q_i' := \hat{r}_{D'}(x, y_i')$ and $\delta_i := q_i - q_i' \in [-\Delta_r, \Delta_r]$. Define

$$A_i(q, \delta) := \mathbb{E}_u\left[\Phi\left(\frac{\tilde{\lambda} + \beta M u - q - \delta}{\sigma_Z}\right)\right], \qquad B_i(q, \delta) := \mathbb{E}_u\left[\Phi\left(\frac{q + \delta - \tilde{\lambda} - \beta M u}{\sigma_Z}\right)\right].$$

Then $1 - p_i(D) = A_i(q_i', \delta_i)$ and $p_i(D) = B_i(q_i', \delta_i)$, and similarly $1 - p_i(D') = A_i(q_i', 0)$, $p_i(D') = B_i(q_i', 0)$. The ratio of halting probabilities is

$$\frac{\Pr[t^* = t \mid D]}{\Pr[t^* = t \mid D']} = \prod_{i=1}^{t-1} \frac{A_i(q_i', \delta_i)}{A_i(q_i', 0)} \cdot \frac{B_t(q_t', \delta_t)}{B_t(q_t', 0)}.$$

Worst-case shifts. The function $A_i(q_i', \delta)$ is decreasing in δ (larger δ shrinks Φ's argument), so $A_i(q_i', \delta_i)/A_i(q_i', 0)$ is maximized at $\delta_i = -\Delta_r$ for $i < t$. The function $B_t(q_t', \delta)$ is increasing in δ, so $B_t(q_t', \delta_t)/B_t(q_t', 0)$ is maximized at $\delta_t = +\Delta_r$.

Worst-case query values. After fixing the shifts, each ratio factor depends only on q_i' through the arguments of Φ. It holds that (see Lemma C.2)

(i) For $i < t$: $A_i(q_i', -\Delta_r)/A_i(q_i', 0)$ is non-decreasing in $q_i' \in [0, R_{\max}]$, maximized at $q_i' = R_{\max}$.

(ii) For $i = t$: $B_t(q_t', \Delta_r)/B_t(q_t', 0)$ is non-increasing in q_t', maximized at $q_t' = 0$.

Final bound. Substituting $q_i' = R_{\max}$ for $i < t$ (with shift $\delta_i = -\Delta_r$) and $q_t' = 0$ (with shift $\delta_t = +\Delta_r$):

$$\frac{A_i(R_{\max}, -\Delta_r)}{A_i(R_{\max}, 0)} = \frac{\mathbb{E}_u[\Phi((\tilde{\lambda} + \beta M u - R_{\max} + \Delta_r)/\sigma_Z)]}{\mathbb{E}_u[\Phi((\tilde{\lambda} + \beta M u - R_{\max})/\sigma_Z)]},$$
$$\frac{B_t(0, \Delta_r)}{B_t(0, 0)} = \frac{\mathbb{E}_u[\Phi((\Delta_r - \tilde{\lambda} - \beta M u)/\sigma_Z)]}{\mathbb{E}_u[\Phi((-\tilde{\lambda} - \beta M u)/\sigma_Z)]}.$$

The factors for $i = 1, \ldots, t-1$ are identical (since the expression depends on i only through u_i, and the u_i's are i.i.d.), so their product is the first factor raised to the power $t - 1$. Taking logarithms yields (4.1).

Composition. By basic composition of DP mechanisms, PrivITP satisfies $(\varepsilon_1 + \varepsilon_2^{\text{post}}(t), \delta)$-DP ex-post. □

Lemma C.2 (Monotonicity of Gaussian CDF ratios under uniform averaging)**.** *Let $\tau \sim \text{Unif}[Q, R]$ with $Q < R$, and let $\Delta > 0$. Define, for $x \in \mathbb{R}$,*

$$R_1(x) := \frac{\mathbb{E}\big[\Phi(\tau - x + \Delta)\big]}{\mathbb{E}\big[\Phi(\tau - x)\big]}, \qquad R_2(x) := \frac{\mathbb{E}\big[\Phi(x - \tau + \Delta)\big]}{\mathbb{E}\big[\Phi(x - \tau)\big]}.$$

Then $R_1(x)$ is non-decreasing in x and $R_2(x)$ is non-increasing in x.

Proof. The sign condition for both ratios reduces to log-concavity of a single auxiliary function, which is then established via the Prékopa–Leindler inequality.

Write $N(x) = \int_Q^R \Phi(\tau - x + \Delta)\, d\tau$ and $D(x) = \int_Q^R \Phi(\tau - x)\, d\tau$, so that $R_1(x) = N(x)/D(x)$. Since $\frac{d}{dx}\Phi(\tau - x + \Delta) = -\phi(\tau - x + \Delta)$,

$$R_1'(x) = \frac{N'(x)\, D(x) - N(x)\, D'(x)}{D(x)^2} = \frac{-\int_Q^R \phi(\tau - x + \Delta)\, d\tau \cdot D(x) \;+\; N(x) \cdot \int_Q^R \phi(\tau - x)\, d\tau}{D(x)^2}.$$

Hence $R_1'(x) \geq 0$ if and only if

$$N(x) \int_Q^R \phi(\tau - x)\, d\tau \;\geq\; D(x) \int_Q^R \phi(\tau - x + \Delta)\, d\tau. \tag{15}$$

Substitute $u = \tau - x$ and set $a = Q - x$, $b = R - x$, $L = R - Q$. Define

$$F(t) \;:=\; \int_a^b \Phi(u + t)\, du, \qquad t \in \mathbb{R}.$$

Then

$$N(x) = F(\Delta), \qquad D(x) = F(0), \qquad F'(t) = \int_a^b \phi(u + t)\, du = \Phi(b + t) - \Phi(a + t),$$

so $F'(0) = \int_Q^R \phi(\tau - x)\, d\tau$ and $F'(\Delta) = \int_Q^R \phi(\tau - x + \Delta)\, d\tau$. Inequality (15) therefore reads

$$F(\Delta)\, F'(0) \;\geq\; F(0)\, F'(\Delta), \qquad \text{i.e.,} \qquad \frac{F'(\Delta)}{F(\Delta)} \;\leq\; \frac{F'(0)}{F(0)},$$

which is equivalent to $(\log F)'$ being non-increasing, i.e. *F is log-concave in t*.

Write F as a convolution:

$$F(t) = \int_a^b \Phi(u + t)\, du = \int_{-\infty}^{\infty} \Phi(t - v)\, \mathbf{1}_{[-b,-a]}(v)\, dv = \big(\Phi * \mathbf{1}_{[-b,-a]}\big)(t).$$

Both factors are log-concave: Φ is log-concave and $\mathbf{1}_{[-b,-a]}$ is the indicator of a convex set, hence log-concave. By the Prékopa–Leindler inequality, the convolution of log-concave functions is log-concave. Therefore F is log-concave in t, which establishes $R_1'(x) \geq 0$.

Similarly, set $\widetilde{N}(x) = \int_Q^R \Phi(x - \tau + \Delta)\, d\tau$ and $\widetilde{D}(x) = \int_Q^R \Phi(x - \tau)\, d\tau$, so that $R_2(x) = \widetilde{N}(x)/\widetilde{D}(x)$. Substituting $v = x - \tau$ gives

$$\widetilde{N}(x) = \widetilde{F}(\Delta), \qquad \widetilde{D}(x) = \widetilde{F}(0), \qquad \text{where} \quad \widetilde{F}(t) := \int_{x-R}^{x-Q} \Phi(v + t)\, dv.$$

Since $\frac{d}{dx}\Phi(x - \tau + \Delta) = \phi(x - \tau + \Delta)$, the same computation yields $R_2'(x) \leq 0$ if and only if

$$\widetilde{F}'(\Delta)\, \widetilde{F}(0) \;\geq\; \widetilde{F}(\Delta)\, \widetilde{F}'(0), \qquad \text{i.e.,} \qquad \frac{\widetilde{F}'(\Delta)}{\widetilde{F}(\Delta)} \;\geq\; \frac{\widetilde{F}'(0)}{\widetilde{F}(0)}.$$

But $\widetilde{F}$ is again a convolution of Φ with an indicator, hence log-concave in t by the same argument. Log-concavity gives $(\log \widetilde{F})' = \widetilde{F}'/\widetilde{F}$ non-increasing, which is precisely the required inequality. Therefore $R_2'(x) \leq 0$. □

C.2 Proof of Theorem 4.3 (DP-ITP Regret)

We analyze the regret of PrivITP against an arbitrary comparator policy π^* using a unified smoothed-relu framework: both the Phase 1 threshold noise and the Phase 2 query noise induce policies whose weights differ from Huang et al.'s ideal χ^2-policy by a bounded pointwise amount, and the regret gap is controlled uniformly by a single *smoothed-weight lemma*.

Proof. Define $r(x, y) := \hat{r}_D(x, y)$ for brevity. Let π_χ^β denote the χ^2-regularized policy with clean rewards $r(x, y)$ and empirical threshold λ^*, i.e.,

$$\pi_\chi^\beta(y|x) = \frac{1}{Z^*}\pi_0(y|x) \cdot w^*(x, y), \qquad w^*(y) := \mathrm{relu}\big(\beta^{-1}(r(x, y) - \lambda^*\big),$$

where λ^* is defined by $\frac{1}{n}\sum_{i=1}^n \mathrm{relu}\big(\beta^{-1}(\hat{r}_D(x, y_i) - \lambda)\big) = 1$ (Huang et al., 2025, Lemma C.1). Let π_{PrivITP} denote the actual policy induced by Algorithm 2, marginalizing over all Phase 1 and Phase 2 noise and over the rejection sampling randomness.

Regret Decomposition. Introduce two intermediate policies:

$$\pi_b(y|x) = \frac{1}{Z_b}\pi_0(y|x)\cdot w_b(x,y), \quad w_b(x,y) := \mathrm{relu}\big(\beta^{-1}(r(x,y)-\tilde{\lambda})\big),$$
$$\pi_{\bar{w}}(y|x) = \frac{1}{Z_{\bar{w}}}\pi_0(y|x)\cdot \bar{w}(x,y), \quad \bar{w}(x,y) := \mathbb{E}_{g\sim\mathcal{N}(0,\sigma_Z^2)}\Big[\mathrm{relu}\big(\beta^{-1}(r(x,y)+g-\tilde{\lambda})\big)\Big],$$

where $\tilde{\lambda} = \lambda^* + \zeta$ with $\zeta \sim \mathcal{N}(0,\sigma_X^2)$. The policy π_b uses the noisy threshold $\tilde{\lambda}$ with clean rewards; $\pi_{\bar{w}}$ additionally marginalizes the Phase 2 Gaussian noise.

Now, we decompose the (expected) regret of PrivITP as

$$\mathrm{Reg}(x) = \underbrace{\big[J(\pi^*) - J(\pi_\chi^\beta)\big]}_{\text{(A): skyline}} + \underbrace{\big[J(\pi_\chi^\beta) - J(\pi_b)\big]}_{\text{(B1): Phase 1 noise}} + \underbrace{\big[J(\pi_b) - J(\pi_{\bar{w}})\big]}_{\text{(B2): Phase 2 noise}} + \underbrace{\big[J(\pi_{\bar{w}}) - J(\pi_{\mathrm{PrivITP}})\big]}_{\text{(B3): finite sample}}.$$

Term (A): Skyline. Since π_χ^β is the χ^2-regularized policy with clean $\hat{r}$ and λ^*, Theorem 4.1 of Huang et al. (2025) applies and we get

$$J(\pi^*) - J(\pi_\chi^\beta) \le \sqrt{C^{\pi^*}(x)\varepsilon_{\mathrm{RM}}^2(x)} + \beta\cdot C^{\pi^*}(x) + \beta^{-1}\varepsilon_{\mathrm{RM}}^2(x) \tag{16}$$

for $n = \Omega\Big(\frac{R_{\max}}{\beta}\log\Big(\frac{R_{\max}}{\beta\,\epsilon_{\mathrm{RM}}}\Big)\Big)$.

We now establish a lemma that will help us bound the remaining terms.

Lemma C.3 (Smoothed-weight policy comparison)**.** *Let $w_1, w_2 : \mathcal{Y} \to \mathbb{R}_{\ge 0}$ be non-negative weight functions with normalizations $Z_i = \mathbb{E}_{\pi_{\mathrm{ref}}}[w_i] > 0$, and let $\pi_i(y) = \pi_{\mathrm{ref}}(y)w_i(y)/Z_i$. Suppose $\|w_1 - w_2\|_\infty \le \eta$. Then:*

$$|J(\pi_1) - J(\pi_2)| \le \frac{2R_{\max}\eta}{\max(Z_1, Z_2)}.$$

Proof of Lemma C.3. Observe that

$$\begin{aligned} J(\pi_1) - J(\pi_2) &= \frac{\mathbb{E}_{\pi_{\mathrm{ref}}}[r^* w_1]}{Z_1} - \frac{\mathbb{E}_{\pi_{\mathrm{ref}}}[r^* w_2]}{Z_2} \\ &= \frac{Z_2\mathbb{E}[r^* w_1] - Z_1\mathbb{E}[r^* w_2]}{Z_1 Z_2} \\ &= \frac{Z_2\mathbb{E}[r^*(w_1 - w_2)] + (Z_2 - Z_1)\mathbb{E}[r^* w_2]}{Z_1 Z_2}. \end{aligned}$$

Using $r^* \le R_{\max}$, $\|w_1 - w_2\|_\infty \le \eta$, $|Z_1 - Z_2| \le \eta$, and $\mathbb{E}[r^* w_2] \le R_{\max} Z_2$:

$$|J(\pi_1) - J(\pi_2)| \le \frac{Z_2 R_{\max}\eta + \eta R_{\max} Z_2}{Z_1 Z_2} = \frac{2R_{\max}\eta}{Z_1}.$$

By symmetry, the same bound holds with Z_1 replaced by Z_2, giving the stated result with $\max(Z_1, Z_2)$. □

Term (B1): Phase 1 noise effect. Fix τ. Since relu is 1-Lipschitz, we get

$$\|w^* - w_b\|_\infty = \sup_y \left|\mathrm{relu}\left(\frac{r(x,y)-\lambda^*}{\beta}\right) - \mathrm{relu}\left(\frac{r(x,y)-\lambda^*-\zeta}{\beta}\right)\right| \le \frac{|\zeta|}{\beta}.$$

The normalization $Z^* = 1$ (by definition of λ^*). Thus $|Z_b - Z^*| \le \|w_b - w^*\|_\infty \le |\zeta|/\beta$. Hence, by Lemma C.3 with $\eta = |\zeta|/\beta$, and $\max\{Z^*, Z_b\} \ge Z^* = 1$, we get

$$|J(\pi_\chi^\beta) - J(\pi_b)|\cdot \mathbf{1}_{\mathcal{G}_1} \le \frac{2R_{\max}|\zeta|/\beta}{1} = \frac{2R_{\max}|\zeta|}{\beta}.$$

Taking expectation over ζ, we get

$$|J(\pi_\chi^\beta) - J(\pi_b)| \le \frac{2R_{\max}\mathbb{E}|\zeta|}{\beta} \le \frac{2R_{\max}\sigma_X\sqrt{2/\pi}}{\beta} = O(R_{\max}\sigma_X/\beta). \tag{17}$$

Term (B2): Phase 2 noise effect. Define the Gaussian-smoothed relu $\varphi_\tau : \mathbb{R} \to \mathbb{R}_{\geq 0}$ by

$$\varphi_\tau(z) := \mathbb{E}_{g\sim\mathcal{N}(0,\tau^2)}[\mathrm{relu}(z+g)] = z\Phi(z/\tau) + \tau\phi(z/\tau),$$

where Φ, ϕ are the standard Gaussian CDF and PDF. Setting $\tau = \sigma_Z/\beta$, we have $\bar{w}(x,y) = \varphi_\tau(\beta^{-1}(r(x,y) - \tilde{\lambda}))$.

Note that $\varphi_\tau(z) - \mathrm{relu}(z) \geq 0$ by Jensen inequality applied to the convex relu function. Moreover, observe that

$$\sup_z\big(\varphi_\tau(z) - \mathrm{relu}(z)\big) = \varphi_\tau(0) = \tau\phi(0) = \frac{\tau}{\sqrt{2\pi}},$$

as maximum holds at $z = 0$ by standard calculus on $h(z) = \varphi_\tau(z) - \mathrm{relu}(z)$: $h'(z) = \Phi(z/\tau) > 0$ for $z < 0$ and $h'(z) = \Phi(z/\tau) - 1 < 0$ for $z > 0$. Therefore, we get

$$\|w_b - \bar{w}\|_\infty \leq \frac{\tau}{\sqrt{2\pi}} = \frac{\sigma_Z}{\beta\sqrt{2\pi}}.$$

This yields $|Z_b - Z_{\bar{w}}| \leq \|w_b - \bar{w}\|_\infty \leq \sigma_Z/(\beta\sqrt{2\pi})$.

On the event $\mathcal{G}_1 := \{|\zeta| \leq \beta/2\}$, we have $|Z_b - 1| = |Z_b - Z^*| \leq |\zeta|/\beta \leq 1/2$. Hence, on $\mathcal{G}_1$, we have $Z_b \geq 1/2$. Hence by Lemma C.3 with $\eta = \sigma_Z/(\beta\sqrt{2\pi})$ and $\max(Z_b, Z_{\bar{w}}) \geq Z_b \geq 1/2$, we obtain

$$|J(\pi_b) - J(\pi_{\bar{w}})| \cdot \mathbf{1}_{\mathcal{G}_1} \leq \frac{2R_{\max} \cdot \sigma_Z/(\beta\sqrt{2\pi})}{1/2} = \frac{4R_{\max}\sigma_Z}{\beta\sqrt{2\pi}}.$$

Now $\Pr[\mathcal{G}_1^c] = \Pr[|\zeta| > \beta/2] \leq 2\exp(-\beta^2/(8\sigma_X^2))$, and the value gap is bounded trivially by $R_{\max}$ on $\mathcal{G}_1^c$. Then we get

$$|J(\pi_b) - J(\pi_{\bar{w}})| \leq \frac{4R_{\max}\sigma_Z}{\beta\sqrt{2\pi}} + R_{\max}\Pr[\mathcal{G}_1^c] \leq \frac{4R_{\max}\sigma_Z}{\beta\sqrt{2\pi}} + 2R_{\max}\exp\left(-\frac{\beta^2}{8\sigma_X^2}\right).$$

This yields

$$|J(\pi_b) - J(\pi_{\bar{w}})| \leq O(R_{\max}\sigma_Z/\beta) + 2R_{\max}\exp\left(-\frac{\beta^2}{8\sigma_X^2}\right). \tag{18}$$

Term (B3): Finite-sample and truncation correction. We bound the gap between the Phase 2 target policy $\pi_{\bar{w}}$ and the actual algorithm output π_{PrivITP}, which arises from the finite number of candidates and the handling of unbounded Gaussian query noise via truncation. The analysis proceeds by conditioning on the Phase 1 noise ζ, applying a rejection sampling argument for each fixed ζ, and then averaging.

Conditioning on ζ.

Recall $\tilde{\lambda} = \lambda^* + \zeta$ with $\zeta \sim \mathcal{N}(0, \sigma_X^2)$. Both $\pi_{\bar{w}}$ and π_{PrivITP} depend on ζ through $\tilde{\lambda}$. For each fixed ζ, define the conditional target

$$\pi_{\bar{w}}^{(\zeta)}(y|x) = \frac{\pi_0(y)\,\bar{w}^{(\zeta)}(x,y)}{Z_{\bar{w}}^{(\zeta)}}, \quad \bar{w}^{(\zeta)}(x,y) = \mathbb{E}_{g'}\left[\mathrm{relu}\big(\beta^{-1}(r(x,y)+g'-\tilde{\lambda})\big)\right], \quad Z_{\bar{w}}^{(\zeta)} = \mathbb{E}_{\pi_0}[\bar{w}^{(\zeta)}],$$

and let $\pi_{\mathrm{PrivITP}}^{(\zeta)}$ denote the algorithm's output distribution conditional on ζ (marginalizing over Phase 2 candidates, Phase 2 noise, and rejection sampling uniform samples). The full distributions satisfy $J(\pi_{\bar{w}}) = \mathbb{E}_\zeta[J(\pi_{\bar{w}}^{(\zeta)})]$ and $J(\pi_{\mathrm{PrivITP}}) = \mathbb{E}_\zeta[J(\pi_{\mathrm{PrivITP}}^{(\zeta)})]$. By the triangle inequality

$$|J(\pi_{\bar{w}}) - J(\pi_{\mathrm{PrivITP}})| \leq \mathbb{E}_\zeta\big|J(\pi_{\bar{w}}^{(\zeta)}) - J(\pi_{\mathrm{PrivITP}}^{(\zeta)})\big|. \tag{19}$$

It suffices to bound the conditional gap for each ζ in the "good event" $\mathcal{G}_1 := \{|\zeta| \leq \beta/2\}$. Under this event, $Z_b^{(\zeta)} \geq 1/2$. Since $\varphi_\tau(z) - \mathrm{relu}(z) \geq 0$ (by Jensen applied to the convex relu), we get $\bar{w}^{(\zeta)} \geq w_b^{(\zeta)}$ and thus $Z_{\bar{w}}^{(\zeta)} \geq 1/2$. The gap under $\mathcal{G}_1^c$ can be bounded by $R_{\max}$.

Phase 2 event decomposition (fix $\zeta \in \mathcal{G}_1$).

The Phase 2 algorithm draws n i.i.d. triples $(y_j', g_j', U_j)_{j=1}^n$ where $y_j' \sim \pi_0(\cdot|x)$, $g_j' \sim \mathcal{N}(0, \sigma_Z^2)$, $U_j \sim \mathrm{Unif}[0,1]$. Define the per-trial weight

$$W_j := \mathrm{relu}\big(\beta^{-1}(r(x,y_j') + g_j' - \tilde{\lambda})\big),$$

and truncation level

$$M := (R_{\max} + \sigma_Z L - \tilde{\lambda})/\beta, \qquad L := \sqrt{2\log(n/\delta)},$$

for a parameter $\delta \in (0,1)$ to be chosen.

Note that $\lambda^* \in [-\beta, R_{\max} - \beta]$ from Huang et al. (2025, Lemma C.1). On $\mathcal{G}_1 = \{|\zeta| \le \beta/2\}$ and hence $\tilde{\lambda} \le R_{\max} - \beta/2$, hence $M \ge (\sigma_Z L + \beta/2)/\beta \ge 1/2 > 0$ on $\mathcal{G}_1$.

Now, the algorithm accepts at step j if $U_j \le \min(W_j/M, 1)$. The output Y^* is the first accepted candidate (or a fallback $Y^* \sim \pi_0(\cdot|x)$ if no acceptance occurs). Define three events on the Phase 2 randomness (conditional on Phase 1 randomness ζ):

$$\begin{aligned}
&\mathcal{T} := \{\exists j \in [n] : W_j > M\} \quad \text{(truncation)},\\
&\mathcal{F} := \{\forall j \in [n] : U_j > \min(W_j/M, 1)\} \quad \text{(fallback)},\\
&\mathcal{S} := \mathcal{T}^c \cap \mathcal{F}^c \quad \text{(success: no truncation and at least one acceptance)}.
\end{aligned}$$

Bounding $\Pr[\mathcal{T} \mid \zeta]$.

On $\mathcal{G}_1$, the event $W_j > M$ implies $r(x, y'_j) + g'_j > R_{\max} + \sigma_Z L$. Since $r(x, y'_j) \le R_{\max}$, this implies $g'_j > \sigma_Z L$. By the Gaussian tail bound, for any $L > 0$, we have

$$\Pr[W_j > M] \le \Pr[g'_j > \sigma_Z L] \le e^{-L^2/2}.$$

Substituting $L = \sqrt{2\log(n/\delta)}$, we get $\Pr[g'_j > \sigma_Z L] \le \delta/n$. A union bound over the n trials gives

$$\Pr[\mathcal{T} \mid \zeta] \le n \cdot \delta/n = \delta. \tag{20}$$

Bounding $\Pr[\mathcal{F} \mid \zeta]$.

The per-trial acceptance probability is

$$p_{\text{acc}} := \mathbb{E}_{y'_j, g'_j}[\Pr[U_j \le \min(W_j/M, 1)]] = \mathbb{E}_{y', g'}[\min(W/M, 1)].$$

Note that

$$p_{\text{acc}} \ge \mathbb{E}[(W/M)\mathbf{1}[W \le M]] = \frac{1}{M}\big(Z_{\bar{w}}^{(\zeta)} - \mathbb{E}[W\mathbf{1}[W > M]]\big).$$

We need an upper bound on the second term $\mathbb{E}[W\mathbf{1}[W > M]]$. Since $M > 0$ on the event $\mathcal{G}_1$, the relu argument in W is positive on the event $\{W > M\}$. Furthermore, since $\{W > M\} \subseteq \{g' > \sigma_Z L\}$, we have

$$\mathbb{E}[W\mathbf{1}[W > M]] \le \mathbb{E}[\beta^{-1}(r(x, y') + g' - \tilde{\lambda})\mathbf{1}[g' > \sigma_Z L]].$$

Using $r(x, y') \le R_{\max}$, we get

$$\mathbb{E}[(R_{\max} + g' - \tilde{\lambda})\mathbf{1}[g' > \sigma_Z L]] = (R_{\max} - \tilde{\lambda})\Pr[g' > \sigma_Z L] + \sigma_Z \mathbb{E}\big[g'/\sigma_Z \cdot \mathbf{1}[g'/\sigma_Z > L]\big].$$

Using $R_{\max} - \tilde{\lambda} = \beta M - \sigma_Z L \le \beta M$, $\Pr[g' > \sigma_Z L] \le \delta/n$ and $\mathbb{E}\big[g'/\sigma_Z \cdot \mathbf{1}[g'/\sigma_Z > L]\big] = \phi(L) = e^{-L^2/2}/\sqrt{2\pi} = \delta/(n\sqrt{2\pi})$, we obtain

$$\mathbb{E}[W\mathbf{1}[W > M]] \le \frac{(\beta M + \sigma_Z)\delta}{\beta n} = \frac{(M + \sigma_Z/\beta)\delta}{n}. \tag{21}$$

Since $\lambda^* \ge -\beta$ and on $\zeta \ge -\beta/2$ on $\mathcal{G}_1$, we have $\tilde{\lambda} \ge -3\beta/2$ and hence $M \le (R_{\max} + \sigma_Z L + 3\beta/2)/\beta$ on $\mathcal{G}_1$. Hence $M + \sigma_z/B \le 2M$.

On $\mathcal{G}_1$, we also have $Z_{\bar{w}}^{(\zeta)} \ge 1/2$ and so $p_{\text{acc}} \ge \frac{1}{M}\left(\frac{1}{2} - \frac{2M\delta}{n}\right) = \frac{1}{2M} - \frac{2\delta}{n}$.

Now, by independence of trials, we obtain

$$\begin{aligned}
\Pr[\mathcal{F} \mid \zeta] = \prod_{j=1}^{n}(1 - p_{\text{acc}}) &\le \left(1 - \left(\frac{1}{2M} - \frac{2\delta}{n}\right)\right)^n \le \exp\left(-\frac{n}{2M} + 2\delta\right)\\
&\le \exp\left(-\frac{\beta n}{2R_{\max} + 2\sigma_Z L + 3\beta} + 2\delta\right).
\end{aligned} \tag{22}$$

Output distribution on $\mathcal{S}$.

On $\mathcal{S}$, no trial has $W_j > M$, so the acceptance rule $U_j \le W_j/M$ is the standard rejection sampling rule with a valid envelope M. Define the *truncated* target policy

$$\pi^{(\zeta)}_{\bar{w},M}(y|x) := \frac{\pi_0(y|x)\bar{w}^{(\zeta)}_M(x,y)}{Z^{(\zeta)}_{\bar{w},M}}, \qquad \bar{w}^{(\zeta)}_M(x,y) := \mathbb{E}_g[W(y,g)\mathbf{1}[W(y,g) \le M]],$$

with $W(y,g) := \mathrm{relu}\big(\beta^{-1}(r(x,y) + g_j - \tilde{\lambda})\big)$ and $Z^{(\zeta)}_{\bar{w},M} = \mathbb{E}_{\pi_0}[\bar{w}^{(\zeta)}_M]$.

Lemma C.4. *Conditional on $\mathcal{S}$ (and ζ), the Phase 2 output $Y^* \sim \pi^{(\zeta)}_{\bar{w},M}(\cdot|x)$.*

Proof of claim. For a single trial, the joint density of (y'_j, g'_j) on the event "$W_j \le M$ and accept at j" is

$$p(y,g) = \pi_0(y|x)\phi_{\sigma_Z}(g)\mathbf{1}[W(y,g) \le M] \cdot (W(y,g)/M),$$

with marginal (integrating over g):

$$\int p_j(y,g)\,dg = \frac{\pi_0(y|x)}{M}\int \phi_{\sigma_Z}(g)W(y,g)\mathbf{1}[W(y,g) \le M]\,dg = \frac{\pi_0(y|x)\bar{w}^{(\zeta)}_M(x,y)}{M}.$$

Normalizing, we get $p(y \mid \text{accept at step } j, W_j \le M) = \pi_0(y|x)\bar{w}^{(\zeta)}_M(x,y)/Z^{(\zeta)}_{\bar{w},M} = \pi^{(\zeta)}_{\bar{w},M}(y|x)$. Since each step conditional on acceptance and no truncation gives the same distribution, and the first-accepted-step stopping time does not bias this distribution, the output $Y^* \mid \mathcal{S} \sim \pi^{(\zeta)}_{\bar{w},M}$ exactly. □

Bounding the $\bar{w}$ vs $\bar{w}_M$ gap.

We bound the pointwise difference

$$\begin{aligned}|\bar{w}^{(\zeta)}(x,y) - \bar{w}^{(\zeta)}_M(x,y)| &= \mathbb{E}_g[W(y,g)\mathbf{1}[W(y,g) > M]] \\ &\le \frac{(M + \sigma_Z/\beta)\delta}{n} \le \frac{(R_{\max} + \sigma_Z(L+1) + 3\beta/2)\delta}{\beta n}\end{aligned}$$

by the same computation as in (21) and using $M \le (R_{\max} + \sigma_Z L + 3\beta/2)/\beta$ on $\mathcal{G}_1$. The same bound applies to $|Z^{(\zeta)}_{\bar{w}} - Z^{(\zeta)}_{\bar{w},M}|$. Applying Lemma C.3 with $\eta = \frac{(R_{\max}+\sigma_Z(L+1)+3\beta/2)\delta}{\beta n}$ and since $\max(Z^{(\zeta)}_{\bar{w}}, Z^{(\zeta)}_{\bar{w},M}) \ge Z^{(\zeta)}_{\bar{w}} \ge 1/2$ on $\mathcal{G}_1$, we get

$$|J(\pi^{(\zeta)}_{\bar{w}}) - J(\pi^{(\zeta)}_{\bar{w},M})| \le \frac{2R_{\max}(R_{\max} + \sigma_Z(L+1) + 3/2)\delta}{(\beta n)\cdot(1/2)} = O(R_{\max}(R_{\max} + \sigma_Z L)\delta/\beta n). \tag{23}$$

Assembling the conditional gap.

Note that

$$J(\pi^{(\zeta)}_{\mathrm{PrivITP}}) = \mathbb{E}[r^*(x,Y^*)] = \mathbb{E}[r^*(x,Y^*)\mathbf{1}_{\mathcal{S}}] + \mathbb{E}[r^*(x,Y^*)\mathbf{1}_{\mathcal{S}^c}].$$

We know that $Y^* \mid \mathcal{S} \sim \pi^{(\zeta)}_{\bar{w},M}(\cdot|x)$, so $\mathbb{E}[r^*(Y^*)\mathbf{1}_{\mathcal{S}}] = J(\pi^{(\zeta)}_{\bar{w},M})\Pr[\mathcal{S}]$. Therefore:

$$\begin{aligned}J(\pi^{(\zeta)}_{\bar{w},M}) - J(\pi^{(\zeta)}_{\mathrm{PrivITP}}) &= J(\pi^{(\zeta)}_{\bar{w},M}) - J(\pi^{(\zeta)}_{\bar{w},M})\Pr[\mathcal{S}] - \mathbb{E}[r^*(x,Y^*)\mathbf{1}_{\mathcal{S}^c}] \\ &= J(\pi^{(\zeta)}_{\bar{w},M})\Pr[\mathcal{S}^c] - \mathbb{E}[r^*(Y^*)\mathbf{1}_{\mathcal{S}^c}].\end{aligned}$$

Taking absolute values and using $r^* \in [0, R_{\max}]$, we get $J(\pi^{(\zeta)}_{\bar{w},M}) \in [0, R_{\max}]$ and $\mathbb{E}[r^*(Y^*)\mathbf{1}_{\mathcal{S}^c}] \in [0, R_{\max}\Pr[\mathcal{S}^c]]$ and thus

$$|J(\pi^{(\zeta)}_{\bar{w},M}) - J(\pi^{(\zeta)}_{\mathrm{PrivITP}})| \le 2R_{\max}\Pr[\mathcal{S}^c \mid \zeta].$$

Combining with (23) via triangle inequality

$$|J(\pi^{(\zeta)}_{\bar{w}}) - J(\pi^{(\zeta)}_{\mathrm{PrivITP}})| \le 2R_{\max}\Pr[\mathcal{S}^c \mid \zeta] + O(R_{\max}(R_{\max} + \sigma_Z L)\delta/\beta n).$$

Algorithm 3 PrivITP with Filtered Self-Reporting Composition

Require: Total privacy budget $(\varepsilon_{\text{total}}, \delta)$; stream of prompts $x_1, x_2, \ldots$; PrivITP hyperparameters $(\beta, \sigma_X, \sigma_Z, L)$

```
Set cumulative privacy spend ε_spent ← 0
Compute per-query worst-case bound ε_max from Theorem 4.1 at t = n
for τ = 1, 2, ... do
    if ε_spent + ε_max ≥ ε_total then
        HALT
    end if
    Run PrivITP on x_τ (Algorithm 2), obtaining response y^{*(τ)} and halting time t_τ
    Compute ex-post privacy cost ε_τ ← ε_1 + ε_2^post(t_τ) from Theorem 4.1
    Update ε_spent ← ε_spent + ε_τ
    Release y^{*(τ)}
end for
```

From (20) and (22), we have $\Pr[\mathcal{S}^c \mid \zeta] \le \delta + \exp\left(-\frac{\beta n}{2R_{\max}+2\sigma_Z L+3\beta} + 2\delta\right)$. Setting $\delta \le 1/\sqrt{n}$ and n such that $n/\log n \ge (2R_{\max} + 2\sigma_Z\sqrt{3\log n} + 3\beta)/\beta$, we get $\Pr[\mathcal{S}^c \mid \zeta] \le O(1/\sqrt{n})$.

Therefore, on $\mathcal{G}_1$, we have

$$|J(\pi_{\bar{w}}^{(\zeta)}) - J(\pi_{\text{PrivITP}}^{(\zeta)})| \le O(R_{\max}(R_{\max} + \sigma_Z L)/\beta n^{3/2}) + O(R_{\max}/\sqrt{n}) = O(R_{\max}/\sqrt{n}). \tag{24}$$

Averaging over ζ.

Observe that $\Pr[\mathcal{G}_1^c] = \Pr[|\zeta| > \beta/2] \le 2\exp(-\beta^2/(8\sigma_X^2))$ and $|J(\pi_{\bar{w}}^{(\zeta)}) - J(\pi_{\text{PrivITP}}^{(\zeta)})| \le R_{\max}$ for all ζ. Combining this with (19) and (24), we get

$$\begin{aligned} |J(\pi_{\bar{w}}) - J(\pi_{\text{PrivITP}})| &\le \mathbb{E}_\zeta\left[|J(\pi_{\bar{w}}^{(\zeta)}) - J(\pi_{\text{PrivITP}}^{(\zeta)})| \cdot \mathbf{1}_{\mathcal{G}_1}\right] + R_{\max}\Pr[\mathcal{G}_1^c] \\ &\le O(R_{\max}/\sqrt{n}) + 2R_{\max}\exp(-\beta^2/(8\sigma_X^2)). \end{aligned} \tag{25}$$

Combining (16), (17), (18) and (25), we obtain the regret

$$\text{Reg}(x) \le \sqrt{C^{\pi^*}\varepsilon_{\text{RM}}^2} + \beta C^{\pi^*} + \frac{\varepsilon_{\text{RM}}^2}{\beta} + O\left(R_{\max}\left(\frac{\sigma_X + \sigma_Z}{\beta} + \frac{1}{\sqrt{n}} + \exp\left(-\frac{\beta^2}{8\sigma_X^2}\right)\right)\right).$$

For $\beta \ge 2\sigma_X$, we can subsume the negative exponential term in σ_X/β and obtain

$$\text{Reg}(x) \le \sqrt{C^{\pi^*}\varepsilon_{\text{RM}}^2} + \beta C^{\pi^*} + \frac{\varepsilon_{\text{RM}}^2}{\beta} + \frac{R_{\max}(\sigma_X + \sigma_Z)}{\beta} + \frac{R_{\max}}{\sqrt{n}}.$$

Combining the requirement on n for Term (A) an (B3), the bounds holds for $n = \max\left\{\Omega\left(\frac{R_{\max}}{\beta}\log\left(\frac{R_{\max}}{\beta\,\epsilon_{\text{RM}}}\right)\right), \widetilde{\Omega}\left(\frac{R_{\max}+\sigma_Z}{\beta}\right)\right\}$. □

D Multi-Query Deployment: Proof of proposition 4.5

Proof. Privacy follows directly from Lebensold et al. (2024, Theorem 11): Algorithm 3 is an instance of their Filtered Composition framework with $\varepsilon_{\max,t} = \varepsilon_{\max}$ (worst-case PrivITP bound) and $\varepsilon_{\text{post},t}(o_t) = \varepsilon_1 + \varepsilon_2^{\text{post}}(t_\tau)$ (ex-post PrivITP bound from Theorem 4.1). Since PrivITP is both $(\varepsilon_{\max}, \delta)$-pDP (via its worst-case bound) and $\varepsilon_{\text{post},\tau}$-ex-post-DP (via Theorem 4.1), the filter condition ensures the cumulative ex-post cost never exceeds $\varepsilon_{\text{total}}$, and the pDP bound handles the tail risk in the final query.

The bound is a direct consequence of the stopping rule: the algorithm halts when $\varepsilon_{\text{spent}} + \varepsilon_{\max} \ge \varepsilon_{\text{total}}$, so by Wald's identity applied to the cumulative cost process, $\mathbb{E}[T^*] \cdot \mathbb{E}[\varepsilon_\tau] \ge \varepsilon_{\text{total}} - \varepsilon_{\max}$, giving the stated lower bound (ignoring the boundary term). □

Remark D.1 (Comparison with privacy filters)**.** *An alternative to FSRC is the privacy filter framework of Rogers et al. (2023); Whitehouse et al. (2023), which provides a stopping rule based on the*

cumulative privacy cost computed via advanced composition. The key difference is that privacy filters use the worst-case per-query cost, while FSRC uses the actual ex-post cost. For PrivITP, where the ex-post cost is typically much smaller than the worst-case cost, FSRC yields strictly tighter deployment budgets. Lebensold et al. (2024) demonstrate this advantage empirically, showing improvements in query throughput at the same total privacy budget.

E PrivBoN with Gaussian Noise

Gumbel noise in PrivBoN yields pure ϵ-DP and converges to a closed-form tilted policy. A natural question is whether Gaussian noise—standard in most DP analyses—offers comparable guarantees. We show here that Gaussian noise has qualitatively different properties that make it a less clean match for inference-time alignment, which motivates our choice of Gumbel noise.

Lemma E.1 (Privacy). *Let $\hat{r}_D(x,\cdot) \in [0, R_{\max}]$ with sensitivity Δ_r. Then, Algorithm 1 with $g_i \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$ satisfies **pure** ε-DP with*

$$\varepsilon = \log \frac{\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\Phi\left(z - \frac{R_{\max} - 2\Delta_r}{\sigma}\right)^{n-1}\right]}{\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\Phi\left(z - \frac{R_{\max}}{\sigma}\right)^{n-1}\right]}, \tag{26}$$

where Φ denotes the standard Gaussian CDF.

Proof. PrivBon with Gaussian noise is an instance of Gaussian Report Noisy Max applied to the n reward scores $\hat{r}_D(x,y_1),\ldots,\hat{r}_D(x,y_n)$, each bounded in $[0,R_{\max}]$ with sensitivity Δ_r. The result follows directly from Lebensold et al. (2024, Theorem 8), which establishes pure ε-DP for Gaussian Report Noisy Max over bounded queries with the closed-form expression in (26). The proof identifies the worst-case reward configurations on adjacent datasets $D \sim D'$ via a monotonicity argument, then expresses the privacy loss as a ratio of Gaussian expectations over products of CDFs. □

The bound is computable via numerical integration (see Lebensold et al., 2024, Section 3.4) and depends on the reward range $R_{\max}$, the sensitivity Δ_r, the noise scale σ, and the number of candidates n – but crucially does not require the approximate-DP slack δ. Unlike Gumbel noise, the privacy budget ϵ depends on n through a numerical integral, though the growth is very slow in practice.

Proposition E.2 (Policy). *PrivBoN with Gaussian noise induces the selection probability $w_i = \mathbb{E}_Z[\prod_{j\neq i}\Phi(Z + (\hat{r}_D(x,y_i) - \hat{r}_D(x,y_j))/\sigma)]$, $i \in [n]$, where $Z \sim \mathcal{N}(0,1)$.*

Proof. Define $r_i := \hat{r}_D(x,y_i)$ for brevity. The derivation follows the same symmetry argument as Proposition 3.2. The only difference is the selection probability. Candidate i wins if $r_i + g_i \geq r_j + g_j$ for all $j \neq i$. Conditioning on $g_i = \sigma \cdot z$, where $z \sim \mathcal{N}(0,1)$, we get

$$\Pr(i \text{ wins} \mid g_i = \sigma\cdot z) = \prod_{j\neq i}\Pr(g_j \leq \sigma\cdot z + r_i - r_j) = \prod_{j\neq i}\Phi\left(z + \frac{r_i - r_j}{\sigma}\right).$$

Integrating over z, we have

$$w_i := \Pr(i \text{ wins}) = \int_{-\infty}^{\infty}\varphi(z)\prod_{j\neq i}\Phi\left(z + \frac{r_i - r_j}{\sigma}\right)dz = \mathbb{E}_{Z\sim\mathcal{N}(0,1)}\left[\prod_{j\neq i}\Phi\left(Z + \frac{r_i - r_j}{\sigma}\right)\right].$$

The rest (symmetry reduction, tower property, marginal policy) is identical to the Gumbel case with w_i replacing the softmax weight. □

This probit integral has no closed-form simplification for general n. Unlike the Gumbel case, PrivBON with Gaussian noise does not converge to a tilted policy as $n \to \infty$: Gaussian extreme value theory dominates, and the selection concentrates on the candidate with the highest $\hat{r}_D$, recovering hard BoN in the limit rather than a smoothed distribution.

F Details on Experiments

F.1 Experimental Setup

Compute resources. All experiments were run using 4 parallel 48GB GPU cards and one 140 GB GPU card. The time to generate responses and score them depends on the size of the base and reward models – the larger the models, the longer the inference time.

Datasets. We evaluate on three benchmarks with verifiable ground-truth correctness, assigning $r^\star(x, y) = 1$ for correct answers and 0 otherwise. GSM8K (Cobbe et al., 2021) is the standard 1K-prompt grade-school math test split. MMLU (Hendrycks et al., 2020) covers $\sim$100 college-level multiple-choice questions each in math and chemistry. MATH-500 (Lightman et al., 2023), a widely adopted, representative subset of 500 test problems from the original MATH dataset (Hendrycks et al., 2021). We use zero-shot Chain-of-Thought prompting (Wei et al., 2022) throughout.

Reward models. We use four reward models as the proxy $\hat{r}_{\mathcal{D}}$: Oasst-RM (Pythia-1.4b), Gemma-RM (Gemma-2-2b), Llama-RM (Llama-3-3b) and Armo-RM (Llama-3-8b). The reference policy π_{ref} (denoted interchangeably π_0) is Gemma-2-2B-Instruct and Phi-3-Mini-Instruct.

Sampling and evaluation. For each prompt, we sample 10K responses at temperature 1 from π_{ref}, then draw $M = 50$ bootstrap replicates of n responses and run each algorithm on every replicate. Per-prompt accuracy is the fraction of replicates yielding a correct answer; reported accuracy averages over prompts, with standard errors estimated across prompt distribution.

F.2 Ablation: Noise Scale for PrivBoN

The noise scale σ in PrivBoN controls both the privacy budget ($\varepsilon = 2\Delta_r/\sigma$) and the strength of KL-regularization, with Theorem 3.3 predicting an optimal $\sigma^* = \varepsilon_{\text{RM}}/\sqrt{C^{\pi^*}}$ balancing under-regularization (reward hacking) and over-regularization (suppressing the reward signal). Since C^{π^*} is not computable in practice, to select σ for the experiments above, we sweep σ and report accuracy lift, and expected proxy reward in Figure 2 on a held out set. The resulting profile is single-peaked: small σ allows reward hacking (accuracy drops as PrivBoN approaches BoN's behavior), while a large σ over-regularizes (accuracy drops toward π_{ref}). We use the empirically best σ for the main results in Table 2. The same σ is also taken as a total noise in PrivITP for a fair comparison.

F.3 Ablation: Regularization for ITP

The regularization parameter β in ITP and PrivITP plays a role analogous to σ in PrivBoN, with Theorem 4.3 predicting an optimal β^* that balances χ^2-bias against overoptimization and privacy noise. Sweeping $\beta \in \{0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$, and reporting accuracy lift, and expected proxy reward in Figure 3, we find the empirically best β on a held out set. Crucially, we observe that PrivITP's optimal β is essentially independent of the privacy budget ε—validating Theorem 4.3's decoupling claim that regularization and privacy can be tuned separately. We use the empirically best β for the main results in Table 2.

F.4 Algorithm comparison across reward models and datasets

We compare four algorithms (BoN, PrivBoN, ITP, and PrivITP) across four reward models (Oasst, Gemma, Llama, Armo) and three datasets (GSM8K, MMLU, MATH), fixing the base policy to Gemma-2-2b-Instruct and Phi-3-Mini-Instruct. For each (algorithm, RM, dataset) configuration, we sweep the number of candidates $N \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096\}$ (interchangeably denoted by n). Private algorithms are calibrated to the empirical best σ obtained for PrivBoN from the ablation study; non-private ITP serves as the regret-optimal skyline. We report two quantities per configuration: (a) accuracy lift over π_{ref}, measuring *true-reward* performance, and (b) expected proxy reward $\mathbb{E}[\hat{r}_{\mathcal{D}}]$ under each algorithm's output distribution, measuring how aggressively each algorithm exploits the reward model. Figures 4 to 6 present the results for the base policy Gemma-2-2b-Instruct, one figure per dataset, each with a 4×2 grid: rows index reward models, the left column shows accuracy lift vs. N, the right column shows expected proxy reward vs. N. In a similar way, Figures 7 to 9 present the results for the base policy Phi-3-Mini-Instruct. Figures 10 to 12 compare % accuracy lift of ITP vs. PrivITP for different values of σ on GSM8K, MMLU and

Table 2: Percentage lift in accuracy over base policy Gemma-2-2B-Instruct at $n = 2^{12}$ candidates. Values are mean $\pm$ standard error over all prompts across the GSM8K, MMLU, and MATH test splits. Same (RM, dataset) configuration uses the same noise σ for private variants and the same regularization β for ITP variants.

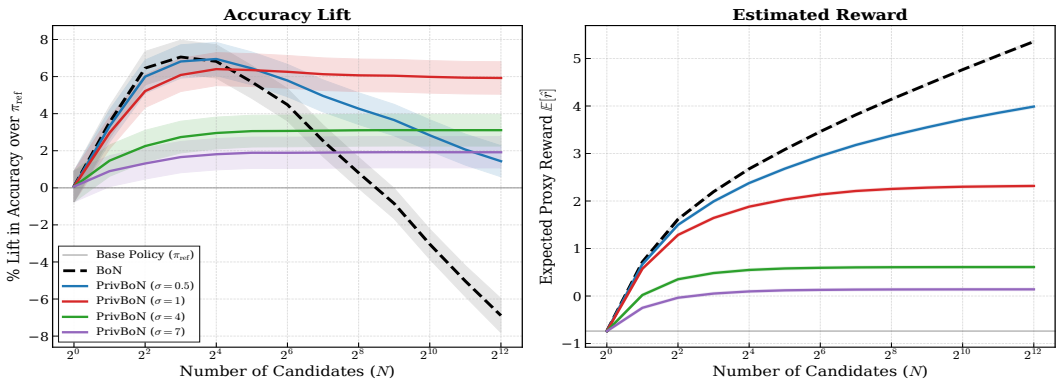
Dataset	Algorithm	Oasst-RM	Gemma-RM	Llama-RM	Armo-RM
GSM8K	BoN	-6.46 ± 0.92	0.75 ± 0.97	16.70 ± 1.12	3.43 ± 1.07
	PrivBoN	5.94 ± 0.88	10.95 ± 0.90	21.46 ± 0.92	7.65 ± 0.87
	PrivITP	6.77 ± 0.90	11.81 ± 0.91	25.67 ± 0.93	14.16 ± 0.90
	ITP	7.36 ± 0.90	12.50 ± 0.92	29.69 ± 0.93	23.16 ± 0.93
MMLU	BoN	10.55 ± 2.21	11.09 ± 2.24	8.59 ± 2.18	23.18 ± 2.89
	PrivBoN	10.14 ± 1.93	9.04 ± 1.71	4.01 ± 1.63	1.12 ± 1.48
	PrivITP	8.26 ± 1.68	14.07 ± 1.93	8.89 ± 1.91	2.49 ± 1.57
	ITP	8.45 ± 1.69	15.44 ± 2.01	9.03 ± 1.99	15.69 ± 2.31
MATH-500	BoN	1.32 ± 1.27	7.38 ± 1.73	20.48 ± 1.86	17.77 ± 1.81
	PrivBoN	-0.48 ± 1.02	3.11 ± 1.33	7.53 ± 1.33	9.34 ± 1.41
	PrivITP	-0.42 ± 1.02	5.44 ± 1.43	14.03 ± 1.53	15.01 ± 1.57
	ITP	0.01 ± 1.05	6.61 ± 1.51	16.56 ± 1.64	16.98 ± 1.68

MATH datasets respectively under Gemma2-2B-Instruct base policy, showing privacy-utility trade-off. Table 2 summarizes the results at $N = 2^{12}$ across all datasets and RMs for Gemma2-2B-Instruct base policy whereas Table 3 summarizes the results at $N = 2^{12}$ across all datasets and RMs for Llama-3.2-3B-Instruct base policy

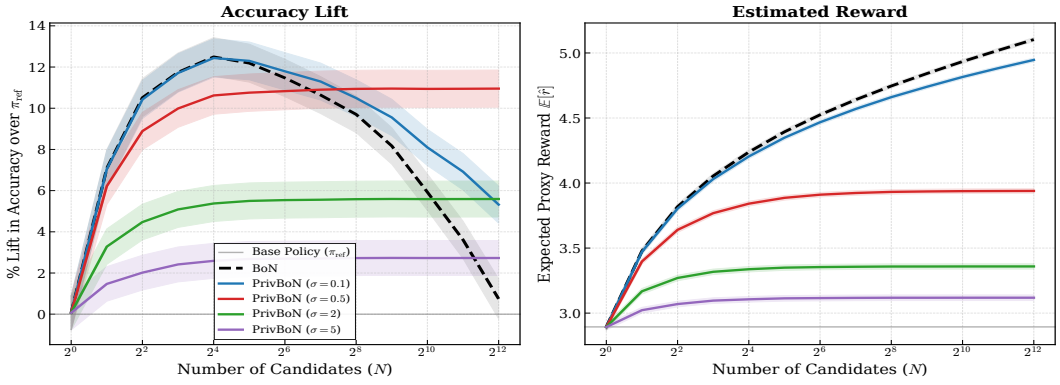
Three patterns hold consistently (except, in some cases, for the harder MATH dataset). *First*, BoN's accuracy lift exhibits the reward-hacking signature: it rises with N initially, then *decreases* starts decreasing, while its expected proxy reward continues to climb monotonically. The two curves diverge—a demonstration of Goodhart's law—confirming that BoN exploits proxy errors at the cost of true reward. *Second*, PrivBoN, ITP, and PrivITP are all scaling-monotone: accuracy lift improves (or plateaus) with N and proxy reward saturates rather than diverging. *Third*, PrivITP dominates PrivBoN on every (RM, dataset) configuration where both provide positive lift, and on GSM8K it recovers between 60% and 95% of the non-private ITP skyline; on MMLU and MATH, where the skyline itself is small or noisy, PrivITP closely tracks ITP within standard error. The proxy-reward panels make the mechanism visible: PrivITP's expected proxy reward saturates at a level above PrivBoN's, indicating that PrivITP's χ^2-pessimism allows more aggressive exploitation of *informative* reward signals while still preventing overfitting to errors. Similar to Huang et al. (2025), we also observe that ITP variants tend to have higher average performance than BoN variants, although in many instances this difference is not statistically significant.

F.5 FSRC COMPOSITION

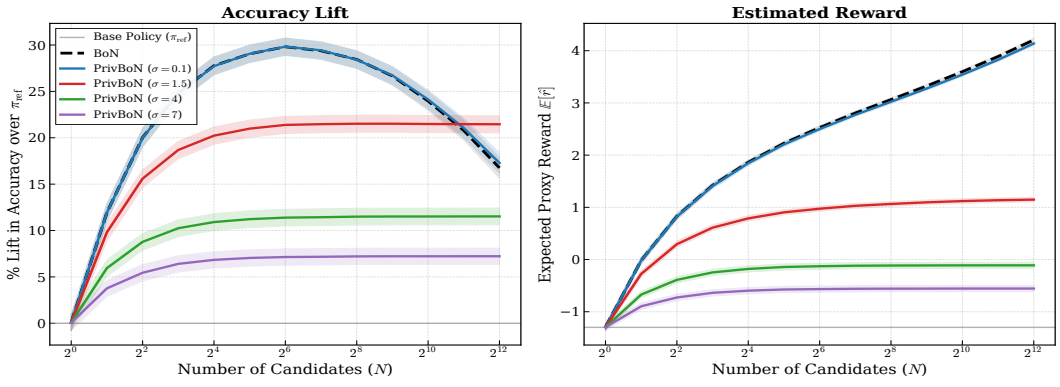
We supplement the main-text composition experiment (Figure 1(bottom right)) with sweeps across reward models using Gemma2-2B-Instruct base policy on GSM8K dataset with $\varepsilon_{\text{total}} = 50$. See Figure 13. Three patterns hold. (i) Absolute query counts increase with σ for both FSRC and standard composition, since larger noise lowers the per-query privacy cost faster than it raises the expected halting time $\mathbb{E}[t]$; both compositors benefit, but FSRC retains its advantage. (ii) The FSRC answers $\sim 3\times$ more queries, reflecting the gap between worst-case $\varepsilon_{\max}$ and realized $\mathbb{E}[\varepsilon_\tau]$. (iii) Empirical halting time admits $t_\tau \ll n$, confirming faster rejection ($\mathbb{E}[t_\tau] = O(1)$). These results confirm that the FSRC gain is structural – driven by the ex-post bound's halting-time dependence.



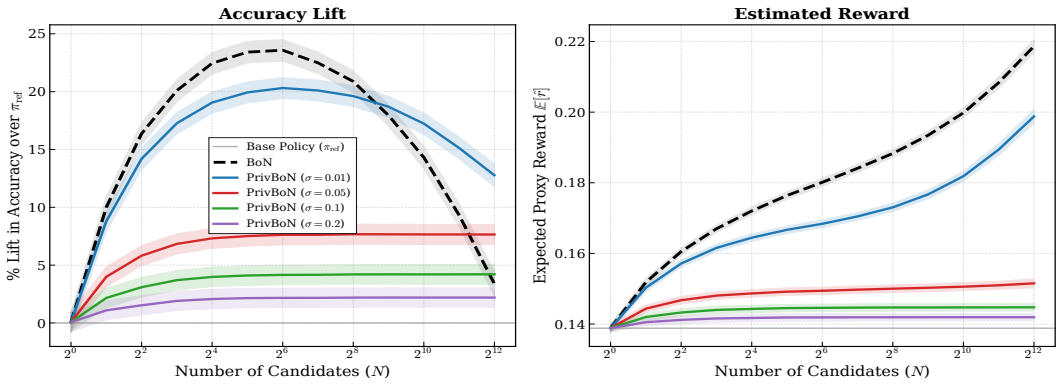
(a) Oasst-RM



(b) Gemma-RM

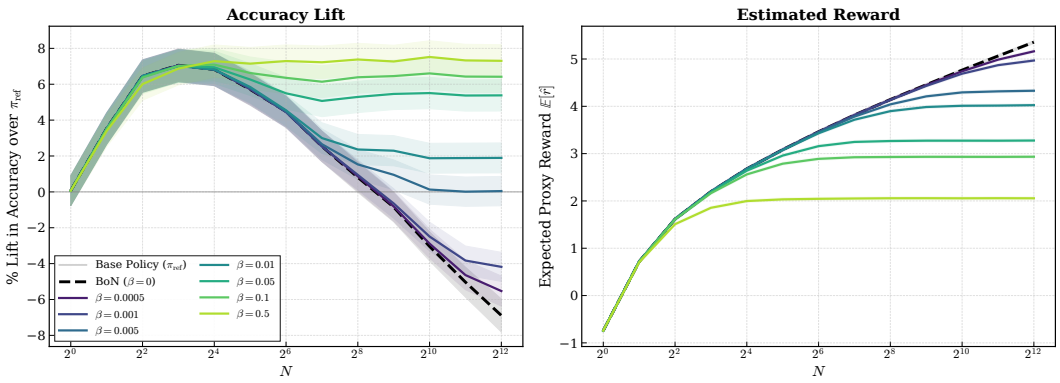


(c) Llama-RM

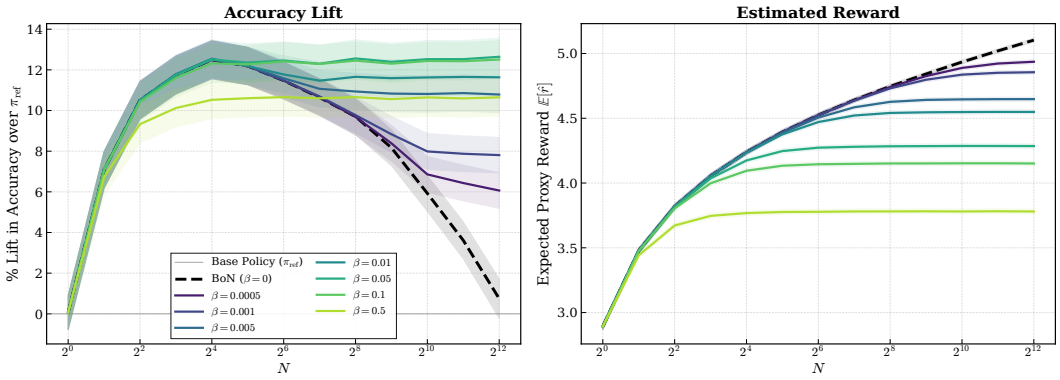


(d) Armo-RM

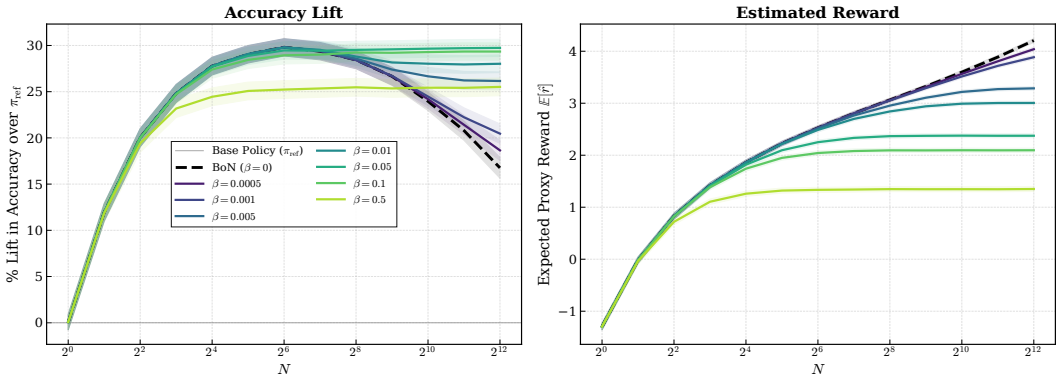
Figure 2: Ablation on noise level σ for PrivBoN on GSM8K dataset with Gemma-2-2b-Instruct as base model



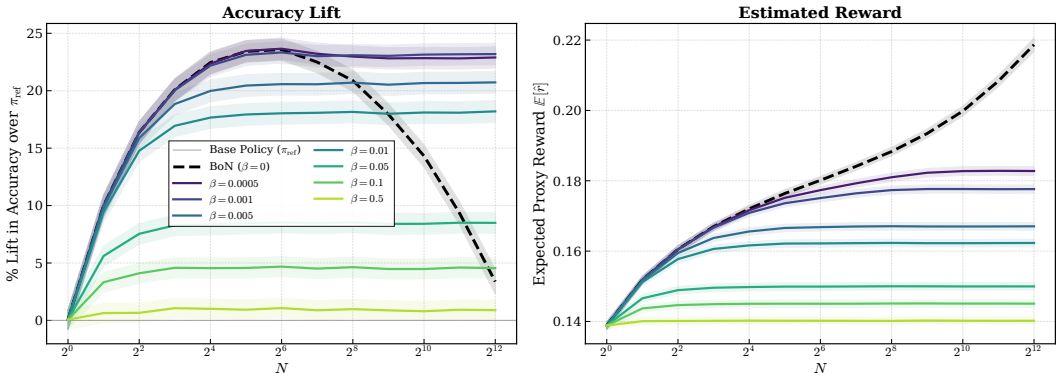
(a) Oasst-RM



(b) Gemma-RM



(c) Llama-RM



(d) Armo-RM

Figure 3: Ablation on regularization β for ITP on GSM8K dataset with Gemma-2-2b-Instruct as base model

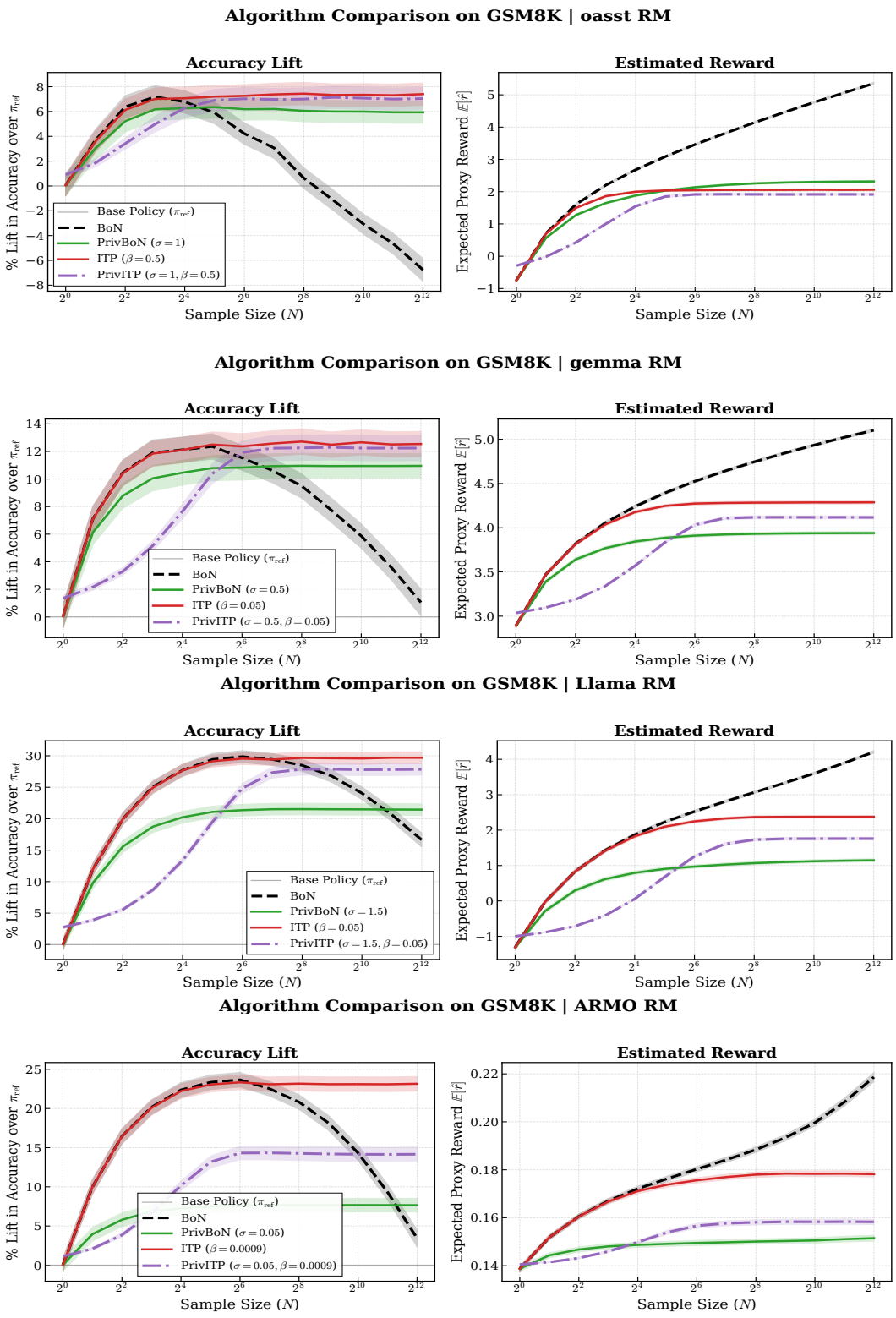


Figure 4: Comparison of BoN, PrivBoN, ITP, and PrivITP in accuracy and estimated reward $\hat{r}$ for GSM8K for four reward models and Gemma-2-2B-Instruct π_{ref}

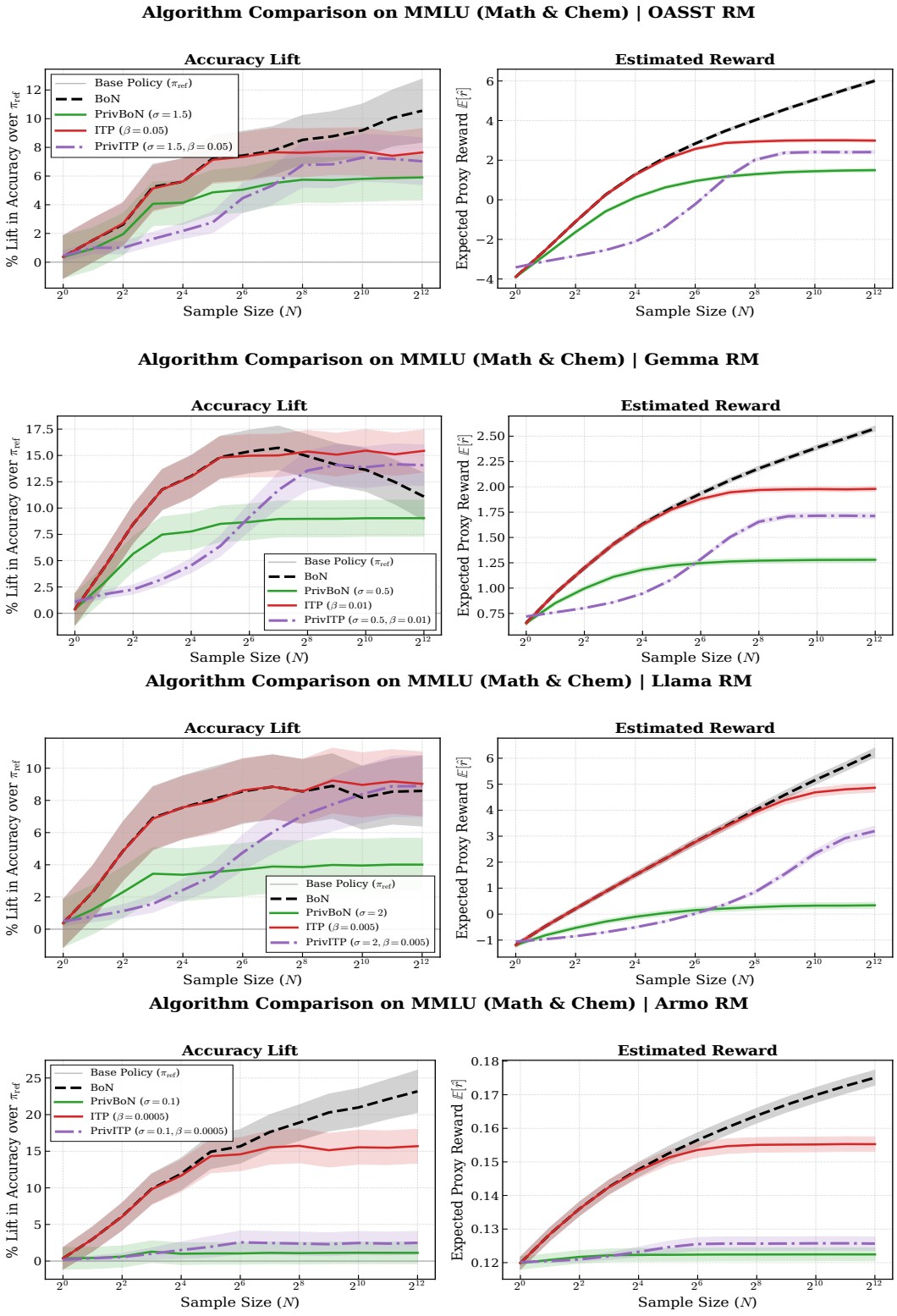
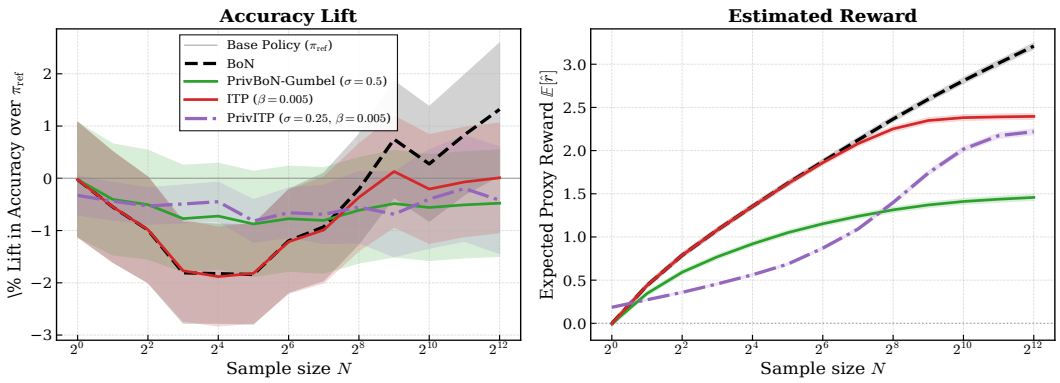
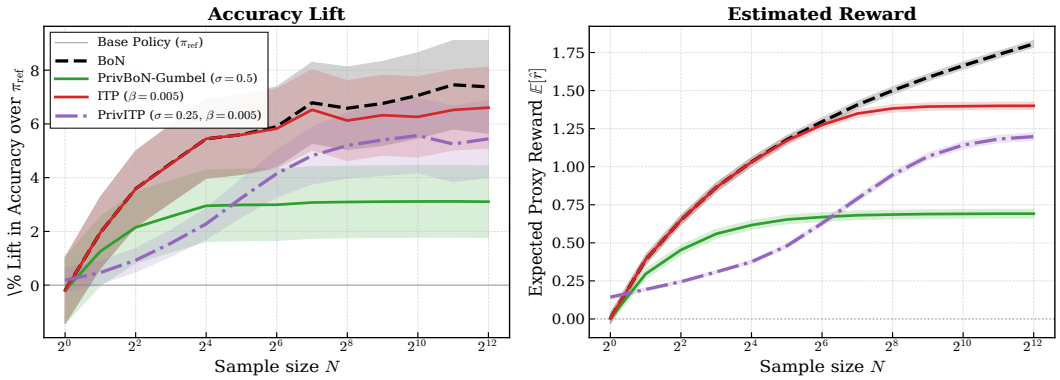


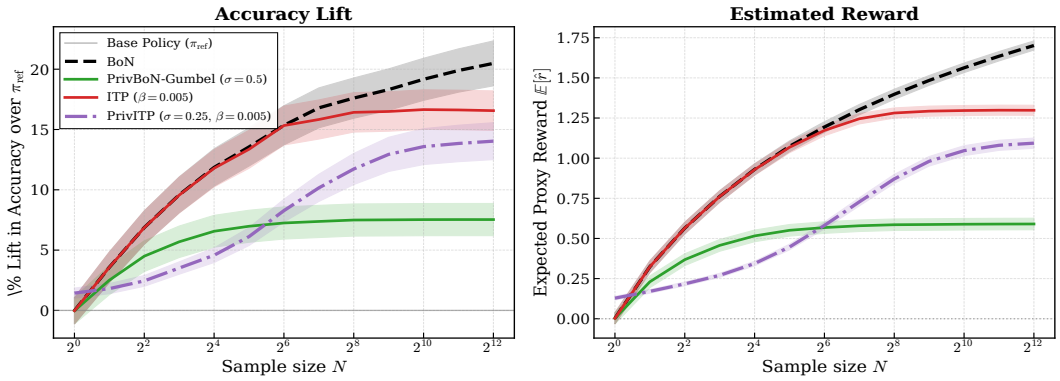
Figure 5: Comparison of BoN, PrivBoN, ITP, and PrivITP in accuracy and estimated reward $\hat{r}$ for MMLU for four reward models and Gemma-2-2B-Instruct π_{ref}



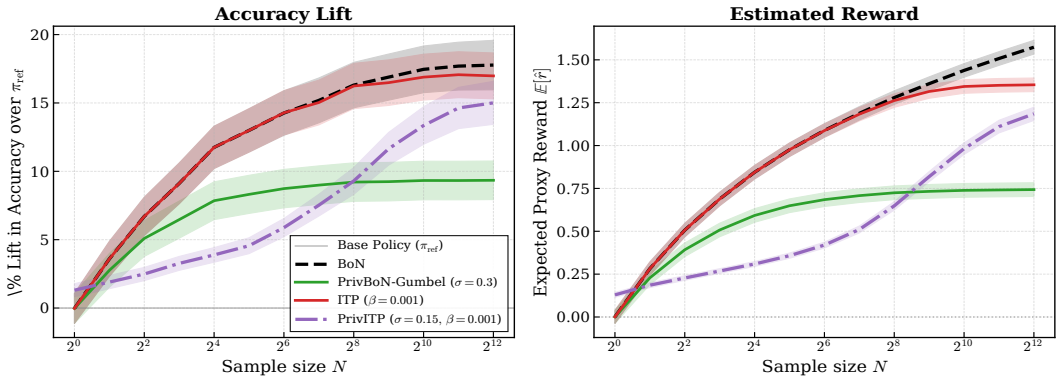
(a) Oasst-RM



(b) Gemma-RM

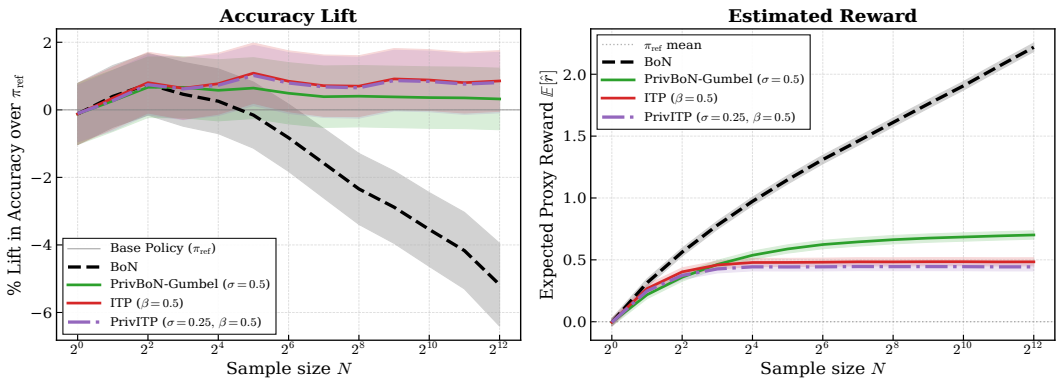


(c) Llama-RM

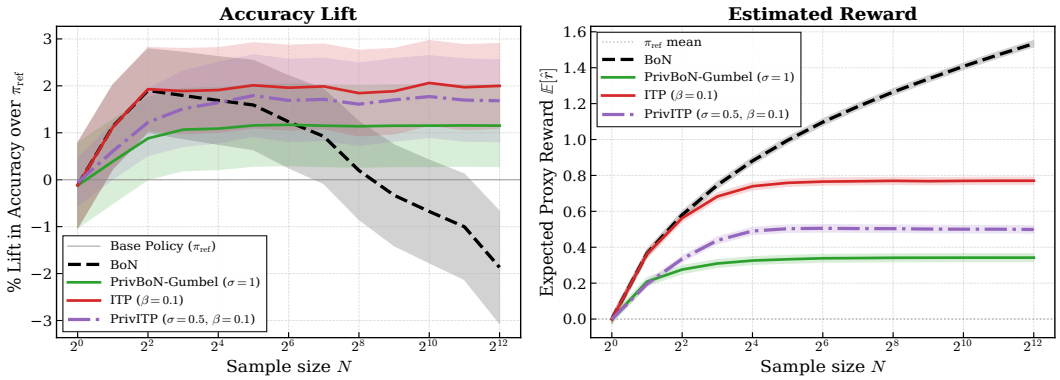


(d) Armo-RM

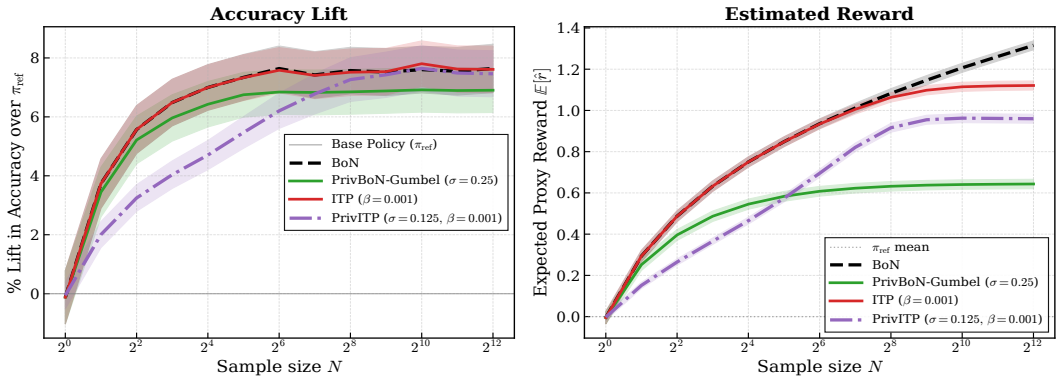
Figure 6: Comparison of BoN, PrivBoN, ITP, and PrivITP in accuracy and estimated reward $\hat{r}$ for MATH for four reward models and Gemma-2-2B-Instruct π_{ref}



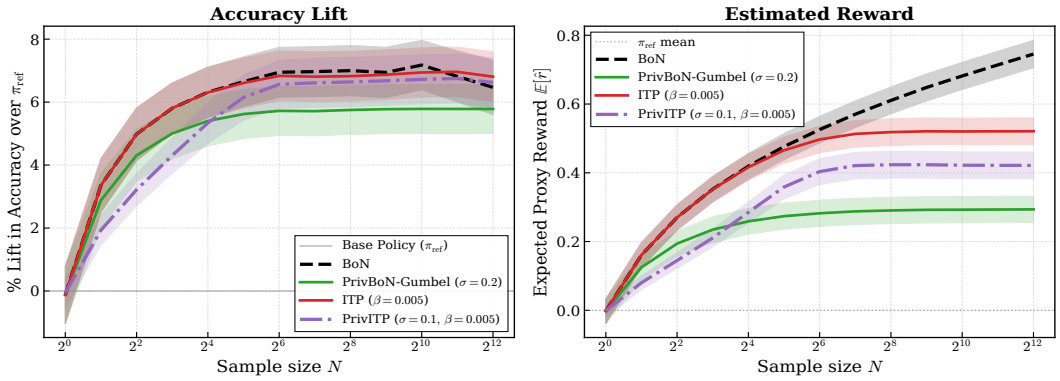
(a) Oasst-RM



(b) Gemma-RM

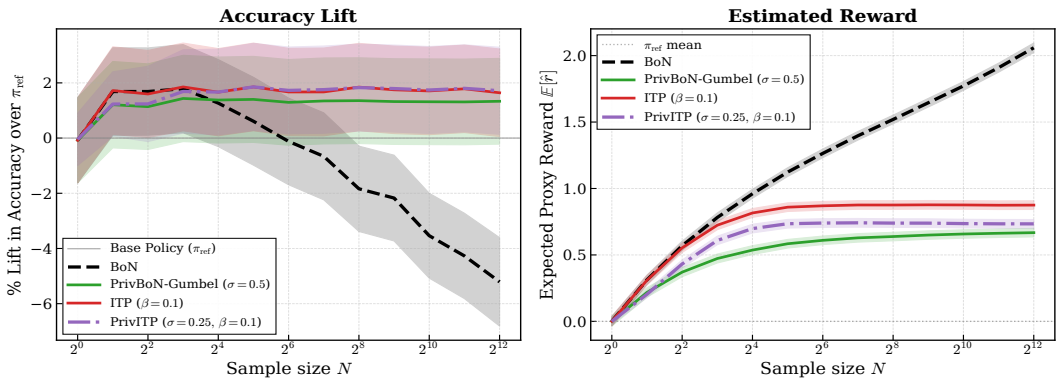


(c) Llama-RM

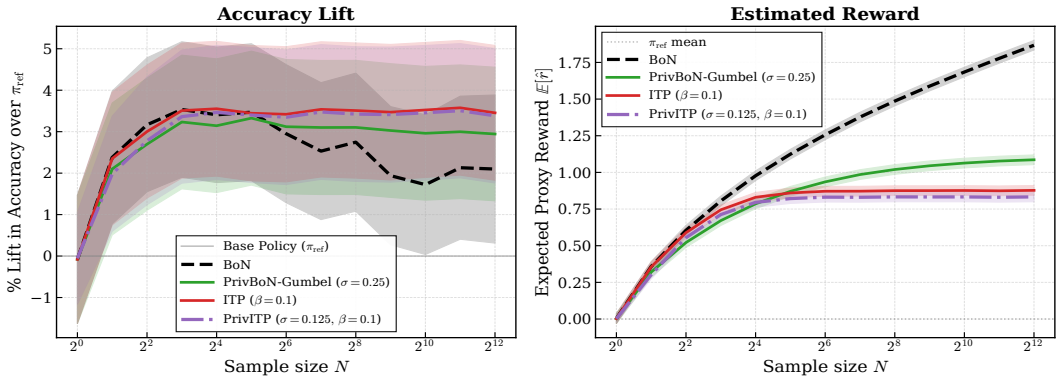


(d) Armo-RM

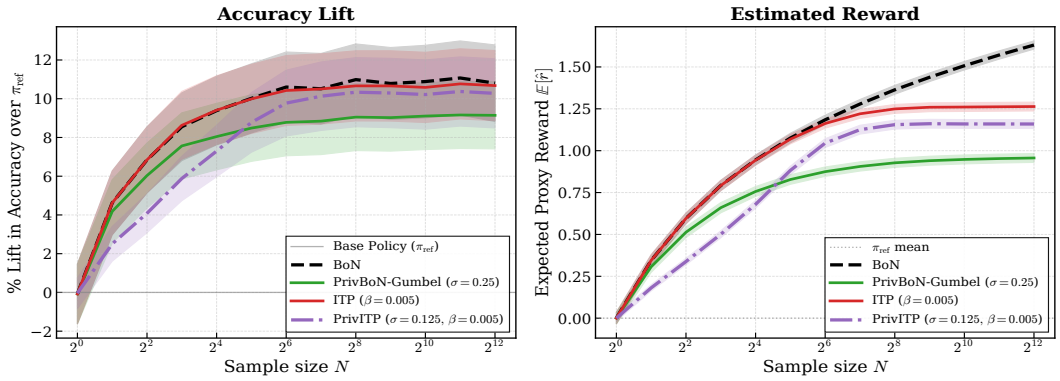
Figure 7: Comparison of BoN, PrivBoN, ITP, and PrivITP in accuracy and estimated reward $\hat{r}$ for GSM8K for four reward models and Phi-3-Mini-Instruct π_{ref}



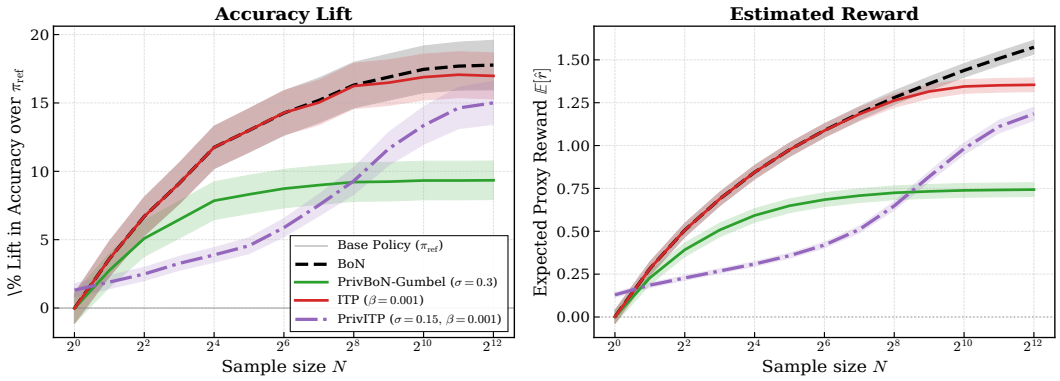
(a) Oasst-RM



(b) Gemma-RM

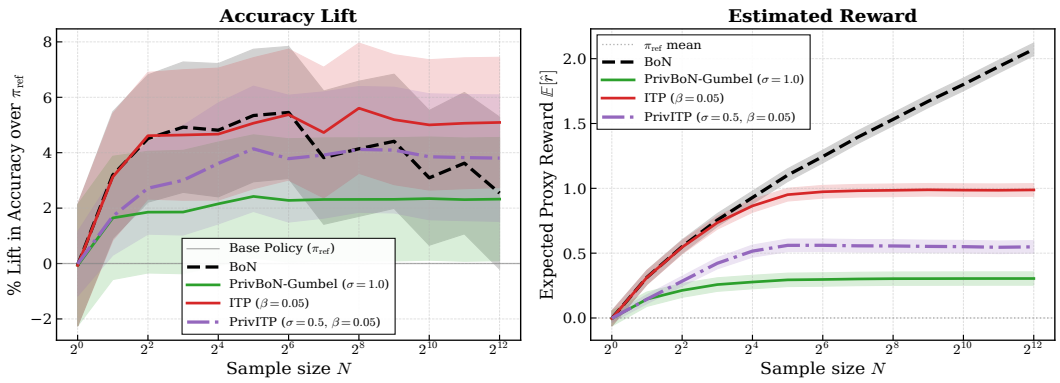


(c) Llama-RM

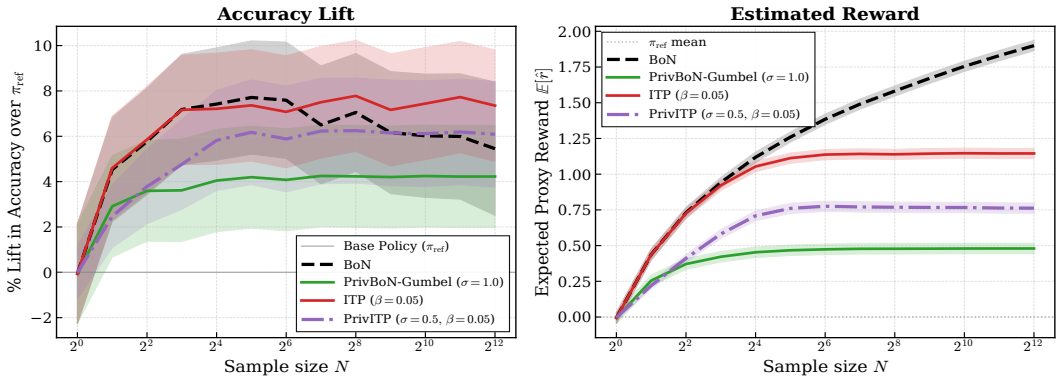


(d) Armo-RM

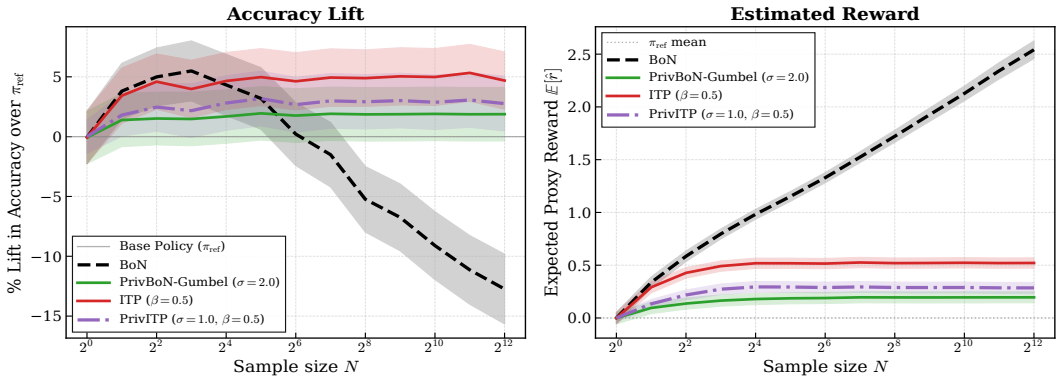
Figure 8: Comparison of BoN, PrivBoN, ITP, and PrivITP in accuracy and estimated reward $\hat{r}$ for MATH for four reward models and Phi-3-Mini-Instruct π_{ref}



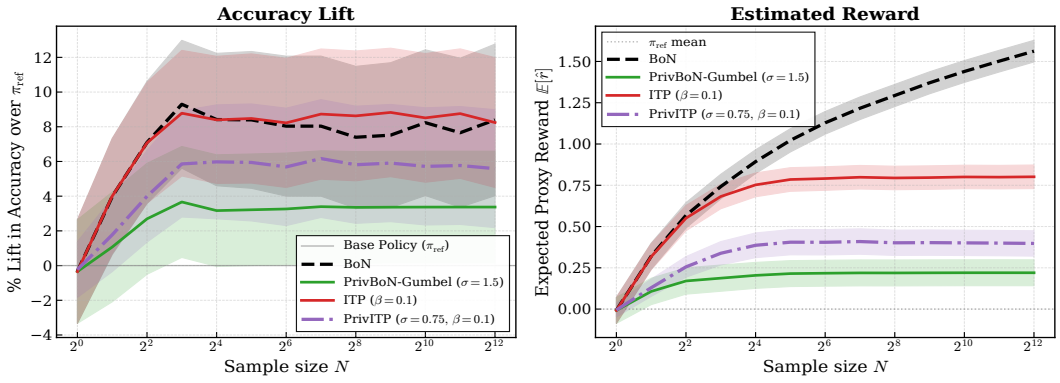
(a) Oasst-RM



(b) Gemma-RM

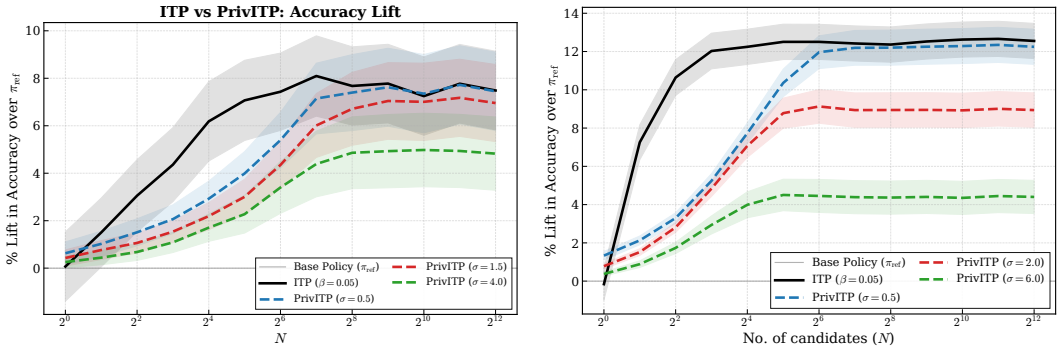


(c) Llama-RM



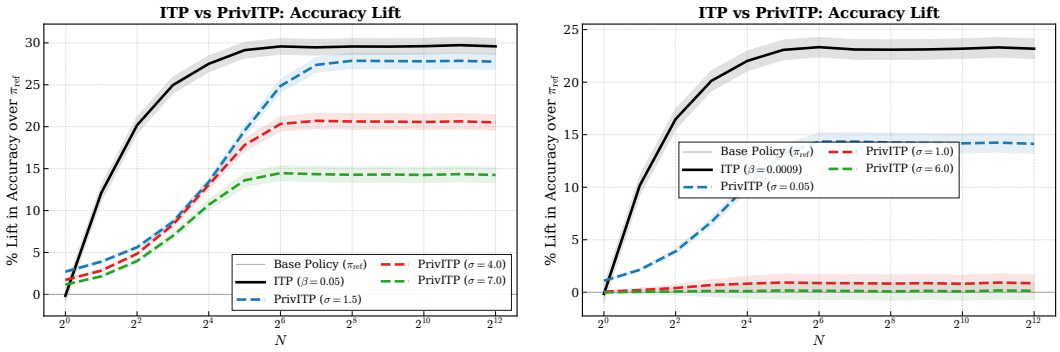
(d) Armo-RM

Figure 9: Comparison of BoN, PrivBoN, ITP, and PrivITP in accuracy and estimated reward $\hat{r}$ for MMLU for four reward models and Phi-3-Mini-Instruct π_{ref}



(a) Oasst-RM

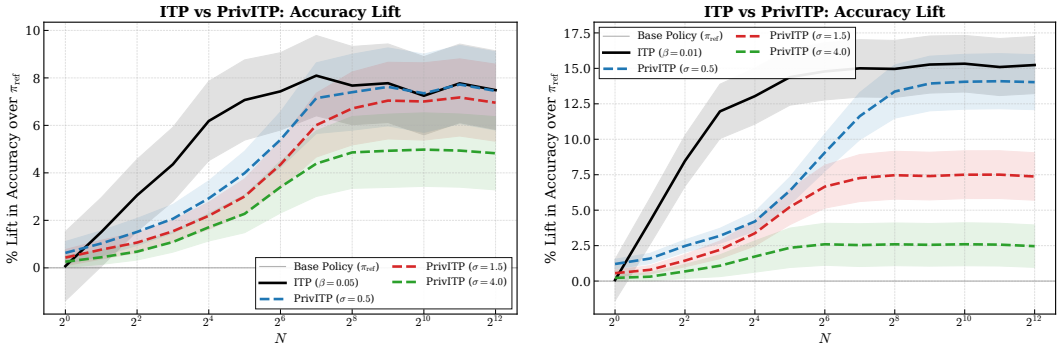
(b) Gemma-RM



(c) Llama-RM

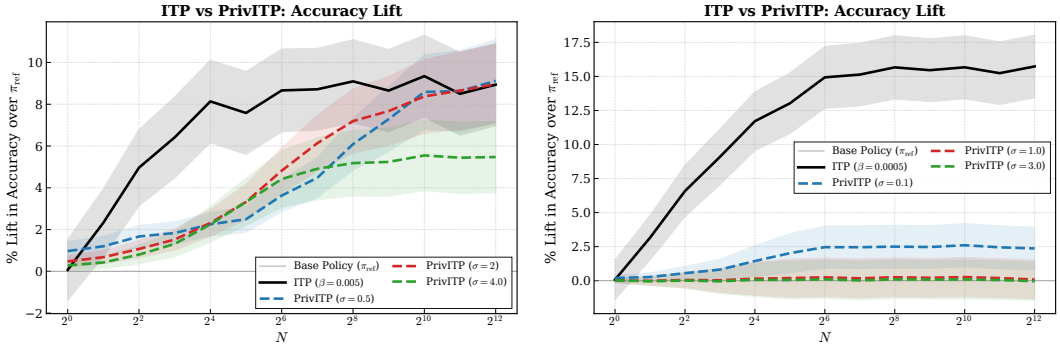
(d) Armo-RM

Figure 10: ITP vs PrivITP for different σ on GSM8K dataset with Gemma-2-2b-Instruct base policy



(a) Oasst-RM

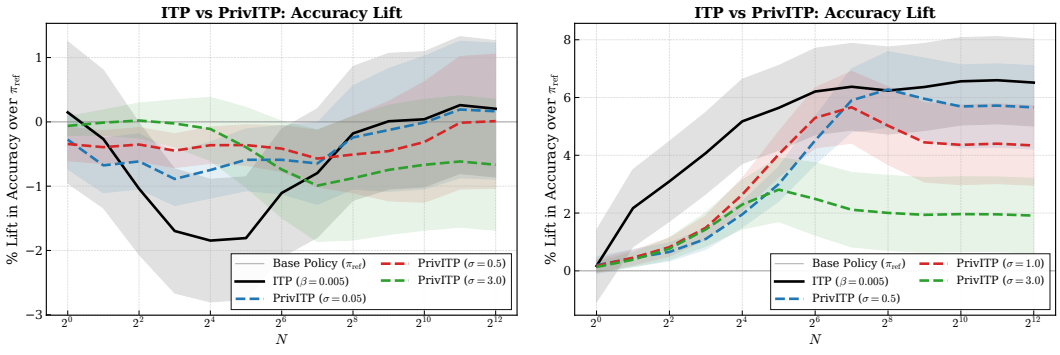
(b) Gemma-RM



(c) Llama-RM

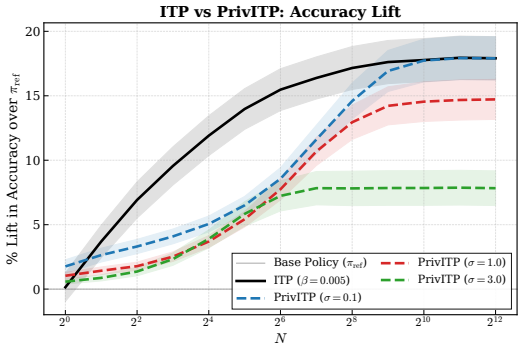
(d) Armo-RM

Figure 11: ITP vs PrivITP for different σ on MMLU dataset with Gemma-2-2b-Instruct base policy

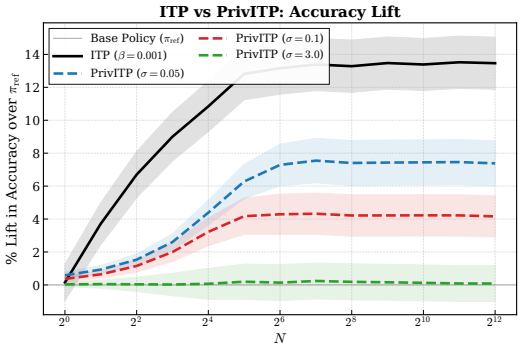


(a) Oasst-RM

(b) Gemma-RM

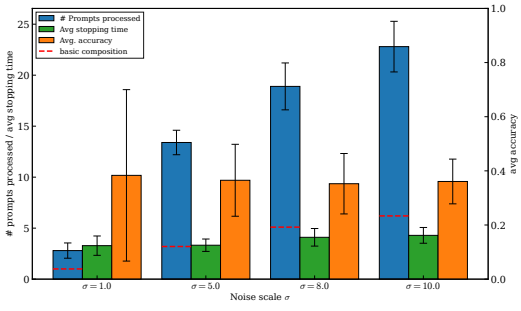


(c) Llama-RM

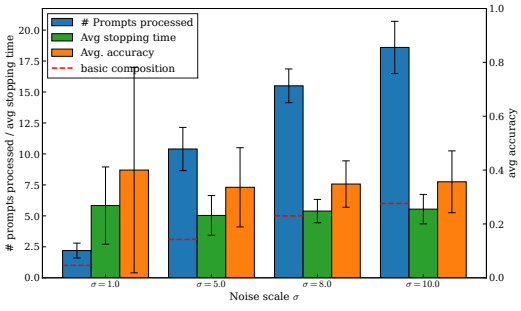


(d) Armo-RM

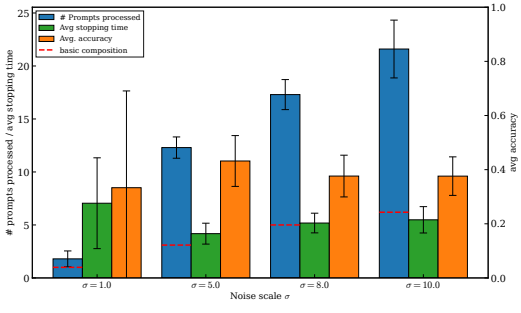
Figure 12: ITP vs PrivITP for different σ on MATH dataset with Gemma-2-2b-Instruct base policy



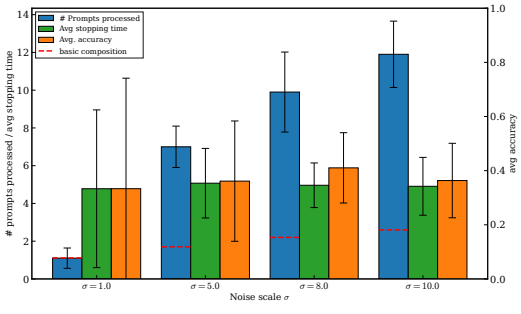
(a) Oasst-RM



(b) Gemma-RM



(c) Llama-RM



(d) Armo-RM

Figure 13: Advantage of FSRC composition in PrivITP's ex-post privacy bound compared to the standard composition in ex-ante bound for different reward models and Gemma-2-2b-Instruct as base model

Table 3: Percentage lift in accuracy over base policy Llama-3.2-3B-Instruct at $n = 2^{12}$ candidates. Values are mean $\pm$ standard error over all prompts across the GSM8K, MMLU, and MATH test splits. Same (RM, dataset) configuration uses the same noise σ for private variants and the same regularization β for ITP variants.

Dataset	Algorithm	Oasst-RM	Gemma-RM	Llama-RM	Armo-RM
GSM8K	BoN	-0.41 ± 1.06	4.68 ± 1.00	19.96 ± 0.75	16.94 ± 0.14
	PrivBoN	3.02 ± 0.84	5.04 ± 0.82	17.71 ± 0.69	11.96 ± 0.11
	PrivITP	4.07 ± 0.85	6.58 ± 0.84	19.63 ± 0.69	14.26 ± 0.11
	ITP	3.98 ± 0.87	6.17 ± 0.87	19.75 ± 0.71	16.31 ± 0.11
MMLU	BoN	7.66 ± 2.68	7.83 ± 2.88	-2.08 ± 2.71	13.80 ± 3.08
	PrivBoN	3.26 ± 1.94	6.70 ± 2.09	2.55 ± 1.97	13.78 ± 3.08
	PrivITP	5.76 ± 2.07	8.78 ± 2.27	4.64 ± 2.10	13.87 ± 2.68
	ITP	7.90 ± 2.25	9.14 ± 2.40	9.27 ± 2.45	8.18 ± 2.28
MATH-500	BoN	-4.48 ± 1.72	3.77 ± 1.87	-8.16 ± 1.82	-3.95 ± 1.86
	PrivBoN	0.34 ± 1.58	2.70 ± 1.62	2.16 ± 1.63	1.35 ± 1.59
	PrivITP	0.46 ± 1.58	4.14 ± 1.66	3.83 ± 1.67	2.21 ± 1.61
	ITP	0.73 ± 1.60	5.79 ± 1.71	6.52 ± 1.80	6.24 ± 1.73

G Additional Experiments and Analysis

G.1 Open-Ended Evaluation (AlpacaEval-2.0)

We evaluate our method on the AlpacaEval-2.0 task (Li et al., 2023) under the standard proxy/gold protocol. In this protocol, a weaker reward model (RM) is used as the proxy $\hat{r}$ for selection, and a strong held-out RM is used as the gold $r^\star$ for scoring. Table 4 reports the gold win-rate (%) versus the base policy, showing the mean and standard error for Gemma-2-2B-Instruct across three proxy RMs (Gemma, Llama, and Armo). We utilize Prometheus-7B-v2.0 (Kim et al., 2024) as the gold $r^\star$ and test across $n \in \{64, 512, 2048, 4096\}$.

Table 4: Gold win-rate (%) vs. the base policy (Gemma-2-2B-Instruct) on AlpacaEval-2.0. We use Prometheus-7B-v2.0 as the gold judge $r^\star$ and evaluate three different proxy reward models.

RM	n	BoN	PrivBoN	ITP	PrivITP
Gemma	64	90.82 ± 1.91	88.35 ± 2.08	90.56 ± 1.99	87.89 ± 2.23
	512	92.19 ± 1.72	89.65 ± 1.97	90.30 ± 1.97	89.78 ± 1.95
	2048	92.06 ± 1.81	89.39 ± 1.98	90.82 ± 1.89	89.97 ± 2.07
	4096	92.38 ± 1.84	90.17 ± 2.02	90.36 ± 1.95	91.15 ± 1.84
Llama	64	92.32 ± 1.87	89.39 ± 1.90	92.32 ± 1.78	89.19 ± 2.03
	512	91.73 ± 1.82	89.32 ± 2.03	91.93 ± 1.66	90.62 ± 1.92
	2048	93.29 ± 1.67	88.87 ± 2.17	92.32 ± 1.73	91.34 ± 1.85
	4096	93.16 ± 1.76	89.39 ± 2.01	92.19 ± 1.73	90.89 ± 2.03
Armo	64	91.47 ± 1.77	88.93 ± 2.10	92.45 ± 1.67	87.63 ± 2.18
	512	92.19 ± 1.64	89.45 ± 2.15	90.49 ± 1.99	90.04 ± 1.89
	2048	91.21 ± 1.94	89.91 ± 1.99	90.17 ± 2.02	89.45 ± 2.08
	4096	92.64 ± 1.75	89.52 ± 2.11	91.02 ± 1.87	89.65 ± 2.03

Observations: First, BoN does not exhibit reward hacking within the tested range, as the gold win-rate remains flat-to-rising as n increases (e.g., Llama increases from 92.32 to 93.16) rather than peaking and declining. This behavior is expected because AlpacaEval win-rates saturate near 90%+, and the proxy RMs track the gold judge closely enough that $n \leq 4096$ does not produce the proxy–gold divergence that drives hacking. Consequently, the private variants match BoN's utility to within standard error at every (n, RM) configuration while adding the privacy guarantee that BoN lacks. The private–non-private gaps are approximately 2–3 points and overlap within one standard error, demonstrating that open-ended generation privacy is obtained at no meaningful utility cost.

G.2 Wall-Clock Overhead and Compute Cost

The actual computational cost for PrivITP is well below a naive $2\times$ overhead because Phase 2 generation is lazy. Rejection sampling returns at the first accepted candidate at index t, meaning candidates only need to be generated on demand. Therefore, Phase 2 costs $\mathbb{E}[t]$ generations rather than n, and experiments show that $\mathbb{E}[t] = O(1)$ for informative RMs. The net generation cost is approximately $n + \mathbb{E}[t]$. Table 5 reports the wall-clock time (seconds per query) of PrivITP compared to PrivBoN at $n = 256$ on the GSM8K dataset using the Phi3-mini-instruct base model on a single H100 GPU.

Table 5: Wall-clock generation time (seconds per query) for PrivBoN and PrivITP at $n = 256$. Evaluated on the GSM8K dataset using the Phi3-mini-instruct base model.

RM	**Oasst**	**Gemma**	**Llama**	**Armo**
PrivBoN	0.4145	0.9003	1.1804	0.4530
PrivITP	0.4704	0.9827	1.2838	0.4812
Overhead	+13.5%	+9.2%	+8.8%	+6.2%

PrivITP's overhead ranges from 6–14% across RMs, which is far below a $2\times$ multiplier. The mechanism overhead itself is negligible in both approaches—PrivBoN draws n Gumbel scalars, while PrivITP draws one scalar Gaussian plus per-round noise, which cost microseconds compared to the n forward passes. Therefore, the gap is dominated primarily by Phase 1's scalar-estimation pass.

G.3 Privacy Budgets and Sensitivity Analysis

In our practical implementation, we bound the sensitivity $\Delta_r = \sup |\hat{r}_D - \hat{r}_{D'}|$ using the maximum reward range $R_{\max}$, since a single example cannot move a bounded-output head beyond its fixed range. This provides an assumption-free bound when reward model training is outside of the deployer's control. Table 6 reports the privacy budgets across different noise levels for GSM8K with Armo-RM, evaluated at $\delta = 10^{-2}$ and $n = 16$. We compute $\varepsilon = 2\Delta_r/\sigma$ for PrivBoN. For PrivITP, the total ex-post cost is $\varepsilon_1 + \varepsilon_2^{\text{post}}$, where Phase 1 costs $\varepsilon_1 = \Delta_r\sqrt{2\ln(1.25/\delta)}/\sigma_X$ and Phase 2 costs $\varepsilon_2^{\text{post}}(\mathbb{E}[t])$ calculated via one-dimensional numerical integration at the mean halting time.

Table 6: Privacy budgets (ε) across varying noise levels (σ) for Armo-RM evaluated on the GSM8K dataset. Budgets are computed at $\delta = 10^{-2}$ and a fallback cap of $n = 16$. PrivITP's ex-post cost ($\varepsilon_1 + \varepsilon_2^{\text{post}}$) is evaluated at the mean halting time $\mathbb{E}[t]$.

σ	ε **(PrivBoN)**	ε_1	$\varepsilon_2^{\text{post}}$	$\varepsilon_1 + \varepsilon_2^{\text{post}}$ **(PrivITP)**
0.10	7.80	24.25	23.25	47.50
0.25	3.12	9.70	8.49	18.19
0.50	1.56	4.85	3.48	8.33
0.75	1.04	3.23	2.22	5.45
1.00	0.78	2.42	1.43	3.85

While PrivBoN spans $\varepsilon \in [0.78, 7.80]$, PrivITP's worst-case ex-post total appears larger because it is evaluated at $\mathbb{E}[t]$ under conservative worst-case query values. However, its realized per-query cost is lower, and FSRC composition exploits this to answer approximately $3\times$ more queries at a fixed total budget. Furthermore, tightening to $\delta = 10^{-5}$ increases ε_1 by a factor of only ≈ 1.56, and $\varepsilon_2^{\text{post}}$ grows at most logarithmically in n because the halting time t is governed by the per-round acceptance probability where $\mathbb{E}[t] = O(1)$.