REVISITING SCALING LAWS FOR REWARD OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling laws for optimization against reward models in AI alignment have pinned down how performance depends on optimization effort—[1]measured by a KL-divergence budget relative to a reference policy. Beyond a certain budget, over-optimization (or reward hacking) can arise: because we optimize against a proxy reward model (distinct from true rewards), performance can plateau or degrade. Naturally, the proxy reward's accuracy depends on how much preference data (often in the form of pairwise comparisons) was used to train it. However, existing research does not cleanly identify how performance jointly scales with the amount of training data and the divergence budget. Our main contribution is to provide an empirically accurate and theoretically grounded scaling law in such context. Performance roughly scales as $\Theta(\sqrt{\min\{\log(M), K\}})$, where M is the number of comparisons in training data and K is the policy's divergence budget. We develop an information-theoretic model to establish this upper bound and prove it is tightly achievable through a constructive procedure. Informed by this, we conduct extensive empirical evaluations using a real-world annotation setup, whereby a large 70B gold reward model generates feedback data and proxy reward models are trained from less capable models (0.6B to 4B). Our scaling law provides an excellent fit (R2 from 97% to 99%), outperforms alternative specifications, and remains robust across model sizes, noise, and optimization procedures (best-of-N or policy tilting). Our evidence suggests that reward optimization is analogous to a surprisingly simple selection task: choosing from a sequence of IID Gaussian random variables using noisy preference feedback.

1 INTRODUCTION

Post-training typically entails learning, either explicitly or implicitly, a proxy 'reward' model of user preferences (Christiano et al., 2017; Ouyang et al., 2022; Rafailov et al., 2023). Such a model remains an imperfect proxy of true user preferences, learned from limited data. As such, it is natural that 'over-optimizing' against such a proxy model may eventually lead to a decrease in the performance gained from post training (Manheim & Garrabrant, 2018; Skalse et al., 2022). A seminal paper by Gao et al. (2023) established empirical scaling laws for this phenomenon. They showed that performance first increased like the square root of the divergence of the post-trained policy from the pre-trained policy, and then beyond some specific divergence budget, performance stalled or decreased. With the advent of long running agents, one only expects such a problem to be exacerbated, potentially leading to catastrophic outcomes (Marklund et al., 2026). This has been the topic of much recent research on superalignment. A consensus recipe that has emerged from the above research is that any post-trained policy should be constrained to within some pre-specified *divergence budget* of a reference policy to prevent over-optimization. While previous literature extensively studies tradeoffs between achievable performance and divergence budget, its dependence on the *amount of preference data* is not well-understood. This raises the natural question:

How should the policy divergence budget be set as a function of the amount of data used to train the proxy reward model?

[1]All dashes in this manuscript are human generated.

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.

It is natural, in particular, that a proxy model trained on limited preference data is more susceptible to over-optimization than one trained with a lot of such data. As such, we seek scaling laws that capture the amount data used to train the proxy reward model. In particular, we would like these scaling laws to inform the performance improvement this data affords us relative to a reference policy. In addition, we wish to characterize how the aforementioned divergence budget ought to be scaled with the amount of preference data. Answering these questions can provide important insights to clarify the fundamental limits of post-training and guide algorithmic scaling recipes in AI alignment.

Contributions. Our work provides a theoretical and empirical characterization of how post-training performance jointly scales with the amount of preference data and the divergence budget.

- *Simple theoretical derivation.* We develop an information-theoretic model that guides our subsequent empirical evaluations. The model yields a tight scaling law, roughly of the form $\Theta(\sqrt{\min\{\log(M), K\}})$, where M is the number of preference comparisons in the training dataset (scale of feedback data) and K is the KL-divergence budget (optimization effort). We give constructive procedures that near-optimally attain this bound.
- *Accurate empirical scaling law.* We provide extensive empirical evidence that our post-training scaling law is highly accurate and provides a statistically better fit than alternative specifications, focusing on the novel predicted dependence $\Theta(\sqrt{\log M})$ on the amount of preference data. We use a real-world annotation setup whereby a large 70B model (serving as the gold reward model) generates preference data while proxy reward models are trained on these data using smaller models. The scaling law is robust across optimization procedures, noisy preferences, and model sizes.
- *Conceptual insight.* Our evidence in support of this scaling law connects post-training to a surprisingly simple selection problem: choosing among a sequence of IID Gaussian random variables from noisy preference data.

2 Related Literature

Alignment and overoptimization. Overoptimization of a misspecified objective function (or reward hacking) is a long-standing concern in AI alignment (Wiener, 1960; Leike et al., 2017). Overoptimization can be defined as a regime in which increasing proxy reward strictly decreases true reward—an extreme form of Goodhart's Law whereby optimization against a certain metric erodes its efficacy (Manheim & Garrabrant, 2018). To counter this, Skalse et al. (2022) define a natural notion of *unhackable rewards* but find that it implies a very restrictive condition: any nontrivial unhackable rewards must induce the same policy ordering. In MDPs, Karwowski et al. (2024) give a geometric explanation of how the proxy and true-optimal policies diverge once optimization pushes the occupancy measure to a face of the polytope where the two rewards disagree. While this strand of literature focuses on mechanisms leading to overoptimization, we instead focus on quantifying its effect on achievable performance. Closer to us, Marklund et al. (2026) study a selection model in which an agent optimizing a consequentialist objective over a large outcome space incurs catastrophic performance with high probability. Our model shares the outcome selection set-up and information-theoretic approach, but we separate policy divergence and preference data budgets and construct policies that achieve matching lower bounds. Our goal is eventually to develop accurate scaling laws.

Post-training scaling laws. Gao et al. (2023) provide the closest empirical antecedent to this work: in a synthetic setup where a fixed gold reward model stands in for human annotators, they measure gold reward as a function of KL-divergence from the base policy and find that it rises like $\sqrt{K}$ and then declines. However, while they vary the size of the preference dataset, they do not identify an explicit scaling law in that dimension.[2] An inverted-U shape also uncovered for direct policy optimization setups without a reward model (Rafailov et al., 2024). Pessimistic adjustments and robust ensembling methods can mitigate the perverse effects of overoptimization (Coste et al., 2024; Eisenstein et al., 2023; Huang et al., 2025). However, previous research often take the proxy model as exogenous and do not ask what specific divergence budget the preference data they are trained on

[2]Quoting the authors, "The scaling of α and β with data size are not as cleanly described as for RM size scaling."

can support. The dependence on $\sqrt{K}$ can be traced back to analyses of selection bias for adaptive data analyses (Russo & Zou, 2016). Related work on AI alignment further characterizes tradeoffs between reward maximization and policy distribution shift (Mroueh, 2024; Laidlaw et al., 2025) under assumptions on the proxy rewards. Still, these works treat the underlying data as exogenous; in our language, the scaling is with respect to K but it does not make explicit the dependence on the amount of preference data M. Instead our work reveals how information obtained from preference data constrains attainable rewards.

Learning from comparisons. The proxy in equation 1 is fit to pairwise preferences under a Bradley–Terry model (Bradley & Terry, 1952; Christiano et al., 2017). Zhu et al. (2023) give the first sample complexity bounds for this pipeline, showing that the MLE converges for linear rewards under Bradley–Terry–Luce and Plackett–Luce, but that a policy trained on the plain MLE can fail while a pessimistic MLE succeeds under a coverage condition. This literature bounds how well M comparisons pin down the reward; we instead ask how much reward gain M comparisons can support at all, irrespective of the estimator. Closest to the combinatorial side of our question is work on selection from noisy comparisons: Falahatgar et al. (2017) establish the comparison complexity of identifying an approximate maximum, and Chen & Suh (2015) that of top-K recovery, with rates governed by the approximation tolerance and the gap at the top-K boundary, respectively. However, these works do not explicitly quantify the corresponding expected reward gain nor they account for divergence constraints.

3 Model and Results

Consider the following setup for alignment. For a prompt–response pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, human preferences are encoded by a *gold* reward $r(x, y)$ that is never observed. In its place we have a proxy $\hat{r}(x, y)$, calibrated on a dataset of M pairwise comparisons: for triples (x, y, y'), an annotator reports which of y, y' is preferred. Given a base policy π_0 that assigns to each prompt a distribution over responses, an *aligned* policy $\hat{\pi}_K$ optimizes the proxy against a divergence budget,

$$\max_{\pi} \ \mathbb{E}_{x,\,\pi}[\hat{r}(x, y)] \quad \text{subject to} \quad \mathbb{E}_x[D_{\mathrm{KL}}(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x))] \ \le \ K. \tag{1}$$

K constrains *optimization pressure*. Because $\hat{r}$ is an imperfect proxy, driving expected proxy reward upward eventually drives expected gold reward downward. Writing $R(K) := \mathbb{E}_{x,\,\hat{\pi}_K}[r(x, y)]$ for the gold reward earned by the aligned policy, Gao et al. (2023) establish an empirical scaling law: $R(K)$ initially behaves like $\sqrt{K}$, and beyond some $K^\star$ is decreasing in K.

The proxy is itself calibrated from M comparisons, so both $R(\cdot)$ and $K^\star$ must depend on M; the scaling law of Gao et al. (2023) does not address how. We seek to answer two related practical questions. First, *how does the scaling law depend on M for a given K*, and in particular, *how should the divergence budget K be set as a function of M?* And second, *what gold reward gain over the base policy does the resulting policy earn?*

3.1 An Information-Theoretic Setup

We propose here a simplified setup that will allow us to make theoretical progress on our questions; candidate scaling laws obtained from the analysis of this setup are then evaluated in the empirical setup of Gao et al. (2023) in §5. For a universe of n potential responses, indexed by i, we associate with each an independent gold reward $R_i \sim \mathcal{N}(\mu, \sigma^2)$. We model a prompt as an *admissible set* $A \subseteq [n]$, with $\Pr(i \in A) = p$ independently across i. Here A is the set of responses that are candidate answers to the realized prompt, and p is the fraction of the universe a prompt renders relevant.[3] We remark that this setup is close to one studied recently by Marklund et al. (2026).

A *preference dataset* contains a label Y_{ij} for each pair (i, j) in a set E of M response pairs. The dataset is *deterministic* when $Y_{ij} = \mathbf{1}\{R_i > R_j\}$, and *random* when the labels are sampled conditionally on the rewards. We write $T := \{(i, j, Y_{ij}) : (i, j) \in E\}$ for the dataset.

Assumption 1 (Information Structure)**.** We assume that $E \perp (R, A)$ so that the pairs in the preference dataset are chosen without future knowledge of the admissible set or information about R. Moreover, $A \perp (R, T)$.

[3] If $A = \varnothing$ we allow a dummy fallback action 0 with deterministic reward $R_0 = \mu$.

We let $\pi_0(\cdot \mid A, E)$ be an arbitrary *reward-blind base policy*. Specifically, an A-valued random variable J distributed according to $\pi_0(\cdot \mid A, E)$ is dependent on A and potentially, E but independent of R. Consequently, $\mathbb{E}_{\pi_0}[R_J] = \mu$. An *informed policy* $\pi(\cdot \mid T, A)$ selects an index $J \in A$ after seeing the preference dataset T and the admissible set A. Define the expected policy divergence and the reward gain by

$$K(\pi; \pi_0) \triangleq \mathbb{E}[D_{\mathrm{KL}}(\pi(\cdot \mid T, A) \,\|\, \pi_0(\cdot \mid A, E))], \qquad \Delta(\pi) \triangleq \mathbb{E}_{\pi}[R_J] - \mu. \tag{2}$$

We next present our results; taken together they provide a theoretical answer to the practical questions posed above.

3.2 A Preview of Our Results

We present here an informal overview of our key result, which holds for both deterministic and random preference datasets:

Theorem (Informal)**.** *Suppose $K(\pi; \pi_0) \le K$ and $|E| \le M$. Then, for any such π,*

$$\Delta(\pi) \lesssim \sigma\sqrt{2\min\{K, \log(pM)\}}.$$

Moreover, there exists a preference dataset E and a resulting proxy $\hat{r}$ for which the optimized policy π satisfies

$$\Delta(\pi) \gtrsim (1 - o(1))\,\sigma\sqrt{2\min\{K, \log(p\sqrt{M})\}}.$$

The asymptotic regime implicit in the statements above requires that $\min\{K, \log(p\sqrt{M})\}$ grow large.[4] That is, the regime is one where both the divergence budget, and the effective number of comparisons for a prompt grows. We briefly discuss the implications of the above result here. First, observe that the lower and upper bounds on $\Delta(\pi)$ are tight: they are within a constant factor of $\sqrt{2}$ of each other. As such, the bounds explain precisely the $\sqrt{K}$ scaling law established by Gao et al. (2023). The bounds also prescribe how to pick the divergence budget K: specifically, we want $K \sim \log(pM)$: any slower and we leave improvement opportunities on the table; any faster and we risk over-optimizing against the proxy. The next section formalizes the above theorem and presents its proof, following which we turn to empirically verifying that the above result holds in a (larger scale) variant of the scaling law study performed by Gao et al. (2023).

4 Scaling laws for reward optimization

4.1 Reward optimization scaling law under deterministic preferences

We now make the preceding bounds precise. We allow the informed policy to use the comparison data directly, so our upper bound applies regardless of how a proxy is fit. For the lower bound, we construct a proxy and a base policy that attain the stated gain. Write $\hat{r}(i; T)$ for the proxy reward assigned to response i using the comparison data T. As in equation 1, the aligned policy $\hat{\pi}_K$ maximizes expected proxy reward within divergence budget K.

We assume for simplicity that $n \ge \sqrt{2M}$ throughout the analysis. This ensures enough possible responses for our construction and leaves the scaling law in M and K unchanged. In practice, it is plausible to assume that the number of possible responses is much larger than our preference dataset.

Theorem 2 (Deterministic preferences)**.** *Under the information structure above, every reward-blind base policy π_0, deterministic preference dataset generated on pairs E with $|E| \le M$, and informed policy with $K(\pi; \pi_0) \le K$ must satisfy*

$$\Delta(\pi) \le \sigma\sqrt{2\min\{K, \log(1 + 2pM)\}}. \tag{3}$$

Reciprocally, there is a universal constant $c > 0$ such that, whenever $p\sqrt{2M} \ge 4$, for every $K \ge 0$ there exist a set of comparison pairs E with $|E| \le M$, a reward-blind base policy, and a proxy such that, for the resulting deterministic preference dataset, every optimizer $\hat{\pi}_K$ satisfies

$$\Delta(\hat{\pi}_K) \ge c\,\sigma\sqrt{\min\{K, \log(p\sqrt{2M})\}}. \tag{4}$$

[4] The dependence on p is not central to our analysis; p is a stylized measure of how much overlap there is in the admissible responses of distinct prompts—a notion of how well preference data generalizes across prompts.

The common upper bound is proved below, the lower-bound construction is given in Appendix A.2.

4.2 THE SAME SCALING LAW HOLDS UNDER NOISY PREFERENCES

So far, we assumed that preferences were deterministic: a response is recorded as preferred whenever it has the larger gold reward. In reality, preference data is noisy. For each $(i, j) \in E$, let $Y_{ij} = 1$ indicate that i is preferred to j. As is standard in RLHF, we model preference labels with the Bradley–Terry model (Bradley & Terry, 1952):

$$\Pr(Y_{ij} = 1 \mid R, E) = \frac{1}{1 + \exp(-(R_i - R_j)/\beta)}, \qquad \beta > 0. \tag{5}$$

Here $\beta > 0$ controls the preference noise: larger β makes labels less sensitive to reward differences. The finite-budget lower bound now depends on β/σ, the noise scale relative to the variation in gold rewards.

Theorem 3 (Noisy preferences)**.** *Suppose preferences follow the Bradley-Terry model with conditionally independent labels and noise scale $\beta > 0$. Under the information structure above, every reward-blind base policy π_0, random preference dataset generated on pairs E with $|E| \le M$, and informed policy with $K(\pi; \pi_0) \le K$ satisfy*

$$\Delta(\pi) \le \sigma\sqrt{2\min\{K, \log(1 + 2pM)\}}. \tag{6}$$

Reciprocally, there is a constant $c(\beta/\sigma) > 0$, depending only on β/σ, such that, whenever $p\sqrt{2M} \ge 4$, for every $K \ge 0$ there exist a set of comparison pairs E with $|E| \le M$, a reward-blind base policy, and a proxy such that, for the resulting random preference dataset, every optimizer $\hat{\pi}_K$ satisfies

$$\Delta(\hat{\pi}_K) \ge c(\beta/\sigma)\,\sigma\sqrt{\min\{K, \log(p\sqrt{2M})\}}. \tag{7}$$

Both lower bounds fix a set of $m = \lfloor\sqrt{2M}\rfloor$ responses and compare each pair once, using $\binom{m}{2} \le M$ comparisons. Deterministic preferences reveal the order of these responses by gold reward. With noisy preferences, we score each response by the fraction of its $m - 1$ comparisons in which it is preferred. Within any nonempty subset of these responses, we then select the response with the highest score. For fixed $\sigma, \beta > 0$, the expected gap between its gold reward and the largest gold reward in that subset tends to zero as m grows. Appendix A.3 gives the full finite-budget argument. We prove the common upper bound below.

Proof. Consider a preference dataset generated on a possibly random pair set E, with $|E| \le M$, and an informed policy π satisfying $K(\pi; \pi_0) \le K$. Both arguments below apply to deterministic preferences and to the Bradley–Terry model.

The divergence budget. Let P_π and P_0 denote the joint laws of (R, E, T, A, J) under π and π_0. These laws differ only in the conditional distribution of the chosen J. The chain rule for relative entropy therefore gives

$$D_{\mathrm{KL}}(P_\pi \| P_0) = \mathbb{E}[D_{\mathrm{KL}}(\pi(\cdot \mid T, A) \,\|\, \pi_0(\cdot \mid A, E))] \le K.$$

Under P_0, the selected index is independent of the rewards. Thus, for every $t > 0$,

$$\mathbb{E}_{P_0}\left[e^{t(R_J - \mu)}\right] = \Pr(A = \varnothing) + \Pr(A \ne \varnothing)e^{t^2\sigma^2/2} \le e^{t^2\sigma^2/2}.$$

The entropy inequality $\mathbb{E}_P f \le D_{\mathrm{KL}}(P\|Q) + \log \mathbb{E}_Q e^f$, applied with $f = t(R_J - \mu)$, gives

$$\Delta(\pi) \le \frac{K}{t} + \frac{t\sigma^2}{2}.$$

Taking the infimum over $t > 0$ gives $\Delta(\pi) \le \sigma\sqrt{2K}$, including $K = 0$ by letting $t \downarrow 0$.

The size of the preference dataset. Let $V(E)$ be the set of responses appearing in E, so $|V(E)| \le 2M$. Conditional on E, the labels depend only on rewards in $V(E)$ and, in the noisy model, independent preference randomness. Every response outside $V(E)$ therefore has conditional mean μ given (T, A). Writing $\pi_i = \pi(i \mid T, A)$ and using the independence of the policy's randomization from R given (T, A), we obtain

$$\Delta(\pi) = \mathbb{E}\left[\sum_{i \in A \cap V(E)} \pi_i (R_i - \mu)\right] \le \mathbb{E}[Z], \qquad Z := \max\big(\{0\} \cup \{R_i - \mu : i \in A \cap V(E)\}\big).$$

For $t > 0$, independence of E, A, and the rewards gives

$$\mathbb{E}\left[e^{tZ}\right] \le 1 + \mathbb{E}\left[\sum_{i \in A \cap V(E)} e^{t(R_i - \mu)}\right] = 1 + p\mathbb{E}|V(E)|e^{t^2\sigma^2/2} \le (1 + 2pM)e^{t^2\sigma^2/2}.$$

Jensen's inequality and minimization over t now imply

$$\Delta(\pi) \le \inf_{t>0} \left\{ \frac{\log(1 + 2pM)}{t} + \frac{t\sigma^2}{2} \right\} = \sigma\sqrt{2\log(1 + 2pM)}.$$

Combining the two bounds proves the upper bounds in Theorems 2 and 3. Notice that neither argument requires the size assumptions used for the lower bounds. □

The constants in the two finite-budget guarantees differ, but both admit the same sharp asymptotic bound.

Corollary 4 (Asymptotic bound on performance)**.** *Fix $\sigma > 0$ and, for Bradley–Terry preferences, $\beta > 0$. Under either preference model, along any sequence (n, M, p, K) with $K \to \infty$ and $p\sqrt{2M} \to \infty$, there exist a pair set E with $|E| \le M$, a reward-blind base policy, and a proxy such that, for the resulting preference dataset, every optimizer $\hat{\pi}_K$ satisfies*

$$\Delta(\hat{\pi}_K) \ge (1 - o(1))\,\sigma\sqrt{2\min\{K, \log(p\sqrt{2M})\}}. \tag{8}$$

5 Experiments

5.1 Summary

In this section, we evaluate our proposed scaling law empirically and test its accuracy against other specifications. We use a real-world annotation set-up and generate preference data from a large gold reward model. We train proxy reward models with 0.6B, 1.7B, and 4B parameters and compare alternative scaling functions. The square-root-logarithmic form fits well across model sizes, preference noise, and optimization methods. It significantly outperforms linear and square-root growth in all tested settings

5.2 Setup

Models and data. Following Gao et al. (2023), we study reward optimization using preferences supplied by a fixed gold reward model, Llama-3.3-Nemotron-70B-Reward. Our proxies are Qwen3 base models with 0.6B, 1.7B, and 4B parameters, each with a scalar reward head. Our training corpus contains 1,000 first-turn prompts from the Arena 140K dataset LMArena (2025), a highly diverse dataset spanning code, mathematics, and other topics (the full mixture is described in Appendix B.1). We sample 600 responses per prompt from the fixed reference policy Qwen3-4B-Instruct at temperature 1.

Preference data. We construct 300 disjoint response pairs per prompt, summing to 300,000 offline comparisons in our largest dataset. We exclude identical-response pairs and duplicate text pairs. In the noiseless setting, the response with highest gold reward is preferred, with exact ties broken uniformly, as in Gao et al. (2023). For the noisy-label experiments, we sample a Bradley-Terry preference with probability $\Pr(y_a \succ y_b \mid x) = \sigma(r(x, y_a) - r(x, y_b))$.

Table 1: **Scaling fits across model size, preference noise, and optimization method.** Each R^2 fits $a + bf(M)$ on all evaluation prompts, bold marks the largest R^2 in each row. The final three columns report Holm-adjusted, two-sided paired t-test p-values against $\sqrt{\log M}$.

			R^2				p-value vs. $\sqrt{\log M}$		
Proxy	Noise	Method	$\sqrt{\log M}$	$\log M$	M	$M^{1/2}$	$\log M$	M	$M^{1/2}$
0.6B	None	BoN	**0.992**	0.990	0.571	0.803	1.000	6.1×10^{-20}	3.4×10^{-9}
0.6B	None	Tilt	**0.991**	0.988	0.564	0.795	1.000	2.7×10^{-20}	3.1×10^{-10}
0.6B	BTL	BoN	0.951	**0.954**	0.537	0.775	0.392	1.8×10^{-20}	7.3×10^{-11}
0.6B	BTL	Tilt	0.947	**0.949**	0.527	0.767	0.708	1.3×10^{-20}	4.7×10^{-11}
1.7B	None	BoN	**0.989**	0.981	0.520	0.757	0.023	8.7×10^{-25}	2.2×10^{-14}
1.7B	None	Tilt	**0.989**	0.982	0.519	0.758	0.025	2.3×10^{-24}	4.3×10^{-15}
1.7B	BTL	BoN	**0.963**	0.957	0.511	0.743	1.000	4.4×10^{-23}	2.0×10^{-10}
1.7B	BTL	Tilt	**0.953**	0.948	0.502	0.734	1.000	1.4×10^{-22}	3.4×10^{-10}
4B	None	BoN	**0.972**	0.964	0.476	0.721	0.003	2.4×10^{-28}	2.7×10^{-17}
4B	None	Tilt	**0.970**	0.963	0.476	0.721	0.002	4.2×10^{-30}	1.5×10^{-18}

Training and comparison budgets. The main scaling experiment increases M from 600 to 300,000 by expanding the set of prompts in a fixed, reward-independent order, while holding the number of comparisons per prompt at 300. We fine-tune all parameters for one pass using the pairwise logistic loss and AdamW at constant learning rate 10^{-5}. Each policy update contains 600 comparisons from two complete prompt groups.

Evaluation. We evaluate on 128 prompts disjoint from training, with 8,192 responses each. Best-of-N selects the highest-proxy-reward response among N candidates. We sweep logarithmically spaced N up to 8,192, corresponding to the conventional optimization-pressure coordinate $K_N = \log N - (N-1)/N$ (Gao et al., 2023). We estimate gold and proxy rewards using the unbiased subset-average estimator of Nakano et al. (2021, Appendix I), which highly reduces the variance of the estimates. Reward gains are measured relative to the mean reward at $N = 1$ over the full response pool for each prompt, and are normalized by a fixed reference gold standard deviation.

5.3 The peak reward follows the predicted square root logarithmic scaling

The comparison-limited regime of Section 3.2 predicts square-root-logarithmic growth of achievable gold-reward gains. We first test this scaling by varying the size of preference dataset M sizes from 600 to $300,000$. For each trained proxy reward (M), we determine the best-of-N response model, varying $N \in \mathcal{G}$ in our evaluated grid bounded by $N \leq 8{,}192$. We tune N as the observed peak of $\widehat{\Delta}_{\max}(M) = \max_{N \in \mathcal{G}} \widehat{\Delta}_M(N)$, where $\widehat{\Delta}_M(N)$ is the mean standardized gold-reward gain.

Figure 1 shows that the peak gold reward closely follows the predicted scaling law. Figure 2 compares the predicted form with affine fits $a + bf(M)$ using the same models and two free parameters for every function. The square-root-logarithmic form gives the best fit, with $R^2 = 0.992$, compared with 0.990, 0.803, and 0.571 for $\log M$, $\sqrt{M}$, and M, respectively.

Table 1 summarizes the fits across proxy sizes, preference noise, and optimization methods, which we examine below.

5.4 Robustness to the optimization algorithm

We test robustness to the optimization algorithm by replacing best-of-N selection with exponential tilting, an idealized form of KL-regularized reinforcement learning that approximates its optimal policy using samples from the reference policy. For inverse temperature $\beta > 0$, the objective $\max_\pi \{\mathbb{E}_\pi[\hat{r}_M] - \beta^{-1} D_{\mathrm{KL}}(\pi \| \pi_0)\}$ has the exponentially tilted optimizer $\pi_{\beta,M}(y \mid x) \propto \pi_0(y \mid x) \exp(\beta \hat{r}_M(x, y))$ (Rafailov et al., 2023). We approximate this policy using responses from the reference policy. The same exponential reweighting is used in soft best-of-N sampling to approximate

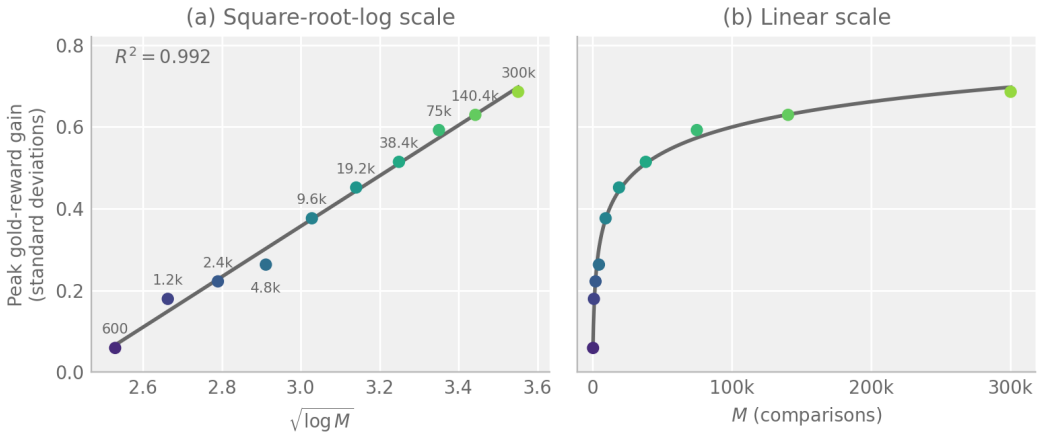


Figure 1: **Gold reward follows square-root-logarithmic comparison-budget scaling.** Ten 0.6B models are selected by uniform spacing in $\sqrt{\log M}$. Each point is the maximum mean gold-reward gain over the evaluated best-of-N grid. Both panels show the same fitted function $a + b\sqrt{\log M}$: left on a square-root-logarithmic axis, right on a linear M axis. Point labels denote comparison counts.

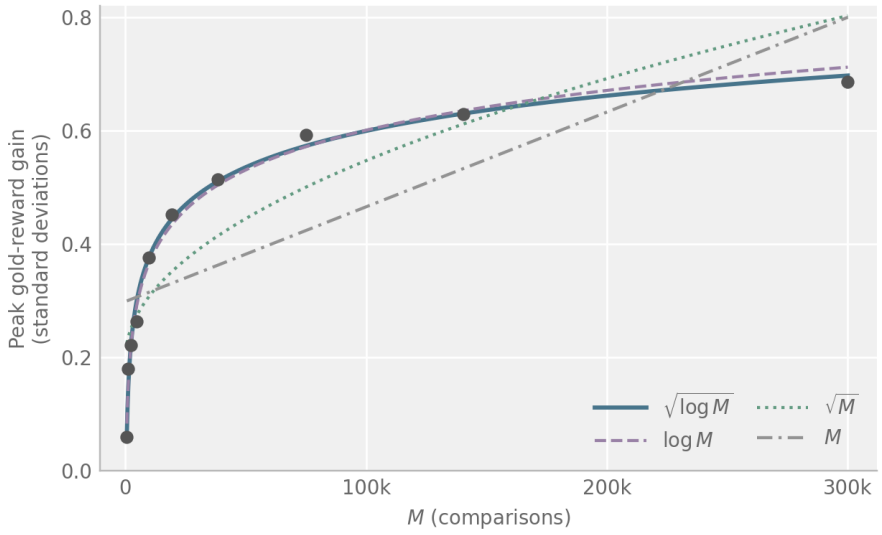


Figure 2: **Comparison of alternative scaling functions.** All curves are fitted to the same ten models as Figure 1. The fits $a + bf(M)$ are shown on a common linear comparison-count axis.

KL-regularized alignment (Verdun et al., 2025), including empirical studies with imperfect reward models (Aminian et al., 2025).

For each evaluation prompt x, we place uniform reference mass on its $S = 8{,}192$ sampled responses and assign the tilted policy weights

$$w_{\beta,M,i}(x) = \frac{\exp(\beta \hat{r}_M(x, y_i))}{\sum_{j=1}^{S} \exp(\beta \hat{r}_M(x, y_j))}. \tag{9}$$

We compute the weighted gold reward and conditional KL exactly on each pool, averaging over the same test prompts and retaining the gold-reward normalization from Section 5.2.

The right column of Figure 5 in Appendix C shows the same square-root-logarithmic scaling under exponential tilting.

5.5 Robustness to preference noise

We next test robustness to preference noise using the Bradley–Terry–Luce (BTL) model (Bradley & Terry, 1952). Labels are drawn once and held fixed throughout training, 33.6% of the 300,000 labels differ from their noiseless counterparts. For each model size, we keep the same model initialization, comparison stream, and evaluation pools, so the comparison isolates the effect of preference noise. Panels (c), (d), (g), and (h) of Figure 5 show that the square-root-logarithmic trend persists under noisy preferences.

The fit is slightly less tight than in the noiseless setting, its agreement with the scaling predicted by our theory is on par with the theory's expected behavior for noisy labels. In particular, the noisy-label theorem predicts the same square-root-logarithmic scaling, but only in a higher-comparison regime. We hypothesize that in practice, additional samples are needed for the noise to average out and for the asymptotic scaling to emerge more cleanly.

5.6 Robustness to proxy model size

Finally, we increase the proxy size from 0.6B to 1.7B and 4B parameters within the Qwen3 base-model family. Both proxies use the same noiseless preferences, comparison order, and evaluation pools.

Table 1 shows that the square-root-logarithmic form remains the best fit among the four specifications for both larger proxies and both optimization methods. For 1.7B, $R^2 = 0.989$ under both methods; for 4B, $R^2 = 0.972$ under best-of-N and 0.970 under tilting. The paired prediction tests also favor this form over $\log M$, $\sqrt{M}$, and M for both larger proxies under both methods, with Holm-adjusted $p < 0.05$ in each comparison.

6 Conclusion

Our work provides a theoretical and empirical characterization of how post-training performance scales jointly with preference data and optimization effort. A simple information-theoretic model yields the scaling $\sqrt{\min(\log(M), K)}$, with constructive procedures that attain this order. Our experiments show that the predicted dependence on preference data provides a strong fit across model sizes, noisy preferences, and optimization procedures. Extending the theory to more general reward distributions and adaptive preference collection, and testing the scaling with human preferences and policy training, are natural directions for future work.

AI use statement

In this work, we used generative AI tools for code, in particular to help debug inference for long training and evaluation runs, and to generate plots. All code was reviewed by the authors. We have not used generative AI tools to help generate synthetic datasets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

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, July 2025. URL `http://arxiv.org/abs/2507.05913`. arXiv:2507.05913 [stat.ML].

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Yuxin Chen and Changho Suh. Spectral MLE: Top-k rank aggregation from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*. PMLR, 2015.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

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

Moein Falahatgar, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Maximum selection and ranking under noisy comparisons. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR, 2017.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023.

Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J. Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Jacek Karwowski, Oliver Hayman, Xingjian Bai, Klaus Kiendlhofer, Charlie Griffin, and Joar Skalse. Goodhart's law in reinforcement learning. In *International Conference on Learning Representations*, 2024.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations*, volume 2025, 2025.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

LMArena. Arena human preference 140k. `https://huggingface.co/datasets/lmarena-ai/arena-human-preference-140k`, 2025. Hugging Face dataset.

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

Henrik Marklund, Alex Infanger, and Benjamin Van Roy. Consequentialist objectives and catastrophe. *arXiv preprint arXiv:2603.15017*, 2026.

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

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, December 2021. URL `https://arxiv.org/abs/2112.09332v3`.

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

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.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pp. 1232–1240. PMLR, 2016.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Claudio Mayrink Verdun, Alex Oesterling, Himabindu Lakkaraju, and Flavio P. Calmon. Soft Best-of-n Sampling for Model Alignment, May 2025. URL http://arxiv.org/abs/2505.03156. arXiv:2505.03156 [cs.IT].

Norbert Wiener. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science*, 131(3410): 1355–1358, 1960.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

A PROOFS

We first establish the upper bound shared by the two preference models. We then prove the lower bound for exact preferences, extend the construction to noisy preferences, and derive the common asymptotic bound. Throughout the lower-bound proofs, $\sigma > 0$. If $\sigma = 0$, every policy has zero gain.

A.1 THE REWARD OPTIMIZATION UPPER BOUND

Consider a preference dataset generated on a possibly random pair set E, with $|E| \le M$, and an informed policy π satisfying $K(\pi;\pi_0) \le K$. Both arguments below apply to exact preferences and to the Bradley–Terry model.

The divergence budget. Let P_π and P_0 denote the joint laws of (R, E, T, A, J) under π and π_0. These laws differ only in the conditional distribution of the chosen J. The chain rule for relative entropy therefore gives

$$D_{\mathrm{KL}}(P_\pi \| P_0) = \mathbb{E}[D_{\mathrm{KL}}(\pi(\cdot \mid T, A) \,\|\, \pi_0(\cdot \mid A, E))] \le K.$$

Under P_0, the selected index is independent of the rewards. Thus, for every $t > 0$,

$$\mathbb{E}_{P_0}\left[e^{t(R_J - \mu)}\right] = \Pr(A = \varnothing) + \Pr(A \neq \varnothing) e^{t^2\sigma^2/2} \le e^{t^2\sigma^2/2}.$$

The entropy inequality $\mathbb{E}_P f \le D_{\mathrm{KL}}(P\|Q) + \log \mathbb{E}_Q e^f$, applied with $f = t(R_J - \mu)$, gives

$$\Delta(\pi) \le \frac{K}{t} + \frac{t\sigma^2}{2}.$$

Taking the infimum over $t > 0$ gives $\Delta(\pi) \le \sigma\sqrt{2K}$, including $K = 0$ by letting $t \downarrow 0$.

The size of the preference dataset. Let $V(E)$ be the set of responses appearing in E, so $|V(E)| \le 2M$. Conditional on E, the labels depend only on rewards in $V(E)$ and, in the noisy model, independent preference randomness. Every response outside $V(E)$ therefore has conditional mean μ given (T, A). Writing $\pi_i = \pi(i \mid T, A)$ and using the independence of the policy's randomization from R given (T, A), we obtain

$$\Delta(\pi) = \mathbb{E}\left[\sum_{i \in A \cap V(E)} \pi_i(R_i - \mu)\right] \le \mathbb{E}Z, \qquad Z := \max\big(\{0\} \cup \{R_i - \mu : i \in A \cap V(E)\}\big).$$

For $t > 0$, independence of E, A, and the rewards gives

$$\mathbb{E}e^{tZ} \le 1 + \mathbb{E}\left[\sum_{i \in A \cap V(E)} e^{t(R_i - \mu)}\right] = 1 + p\mathbb{E}|V(E)|e^{t^2\sigma^2/2} \le (1 + 2pM)e^{t^2\sigma^2/2}.$$

Jensen's inequality and minimization over t now imply

$$\Delta(\pi) \le \inf_{t>0}\left\{\frac{\log(1+2pM)}{t} + \frac{t\sigma^2}{2}\right\} = \sigma\sqrt{2\log(1+2pM)}.$$

Combining the two bounds proves the upper bounds in Theorems 2 and 3. Notice that neither argument requires the size assumptions used for the lower bounds.

A.2 Proof of Theorem 2

It remains to prove the lower bound. We first relate proxy optimization to the construction, then establish its feasibility and reward gain.

The proxy. For either preference model, take $\hat{r}(i; T) = \mathbb{E}[R_i \mid T]$, with $\hat{r}(0; T) = \mu$. Since $A \perp (R, T)$, every informed policy satisfies

$$\mathbb{E}_\pi R_J = \mathbb{E}\left[\sum_i \pi(i \mid T, A)\mathbb{E}[R_i \mid T, A]\right] = \mathbb{E}_\pi \hat{r}(J; T). \tag{10}$$

It therefore suffices to construct a feasible informed policy with the claimed gain. An optimizer exists for the following reason. There are only finitely many possible pairs (T, A): the dataset records pair identities and binary labels, and $A \subseteq [n]$. Ignore pairs of probability zero. A policy is then specified by finitely many choice probabilities. A finite divergence budget requires $\pi(i \mid T, A) = 0$ whenever $\pi_0(i \mid A, E) = 0$. The remaining probabilities are nonnegative and sum to one for each (T, A), so their domain is compact. On this domain, the divergence is continuous: each term $x \log(x/b)$ with $b > 0$ extends continuously to zero by assigning it the value zero. Hence the budget constraint defines a closed subset of this compact domain. The feasible set is nonempty because it contains π_0. Finally, expected proxy reward is a continuous linear function of the choice probabilities, so it must attain its maximum on the feasible set.

The preference dataset and base policy. For the remainder of the proof, write

$$m = \lfloor\sqrt{2M}\rfloor, \qquad q = p\sqrt{2M}, \qquad h = \min\{K, \log q\}.$$

We assumed that $m \le n$ and $q \ge 4$. We then choose a fixed set S of m responses and compare every pair in S. This requires $\binom{m}{2} \le M$ comparisons. Thus the exact preference dataset reveals the full ranking of S.

Let $B = A \cap S$ and $N = |B| \sim \mathrm{Bin}(m, p)$. When $N > 0$, take the base policy to be uniform on B. When $N = 0$, take it to be uniform on A, or to select action 0 if $A = \varnothing$. We will use

$$pm \ge q - 1 \ge 3, \qquad pm \le q, \qquad \log(pm) \ge \tfrac{1}{2}\log q. \tag{11}$$

where the last inequality follows from $q - 1 \ge \sqrt{q}$ for $q \ge 4$.

Selection within the divergence budget. For $K \geq \log 2$, set $s = \min\{N, \lfloor e^K \rfloor\}$. When $N > 0$, we draw an s-element subset U uniformly from B, independently of the rewards and labels conditional on A, and select its highest-ranked element. To bound divergence, consider a reference rule that draws the same subset and then selects uniformly from it. Its marginal distribution on responses is uniform on B, exactly the base policy. Conditional on (T, A, U), selecting one member instead of a uniform member has divergence $\log s$. Marginalizing over U cannot increase relative entropy, so

$$D_{\mathrm{KL}}(\pi(\cdot \mid T, A) \,\|\, \pi_0(\cdot \mid A, E)) \leq \log s \leq K \qquad (N > 0). \tag{12}$$

This feasibility holds for any ranking obtained from the preference dataset.

For $0 \leq K < \log 2$, a subset of size $\lfloor e^K \rfloor$ would give no gain. Instead, when $N \geq 2$, sample a pair uniformly from B and choose its higher-ranked element with probability $(1 + \theta)/2$, where $\theta = \sqrt{K}$, choosing the other member otherwise. Follow the base policy when $N < 2$. The reference rule again chooses uniformly from the same pair. Its marginal is thus the base policy, and the conditional divergence is

$$d(\theta) := \frac{1+\theta}{2} \log(1+\theta) + \frac{1-\theta}{2} \log(1-\theta) \leq \theta^2 = K. \tag{13}$$

The inequality follows from $\log x \leq x - 1$. Marginalizing over the sampled pair proves feasibility, again for any ranking.

Two estimates for the reward gain. For independent standard Gaussians $Z_1, Z_2, \ldots$, define $a_j = \mathbb{E} \max_{1 \leq i \leq j} Z_i$ for $j \geq 1$, and set $a_0 = 0$. We use the Gaussian maximal inequalities:

$$a_j \geq c_0 \sqrt{\log j} \quad (j \geq 2), \qquad a_j \sim \sqrt{2 \log j} \quad (j \to \infty), \tag{14}$$

for a universal $c_0 > 0$.

We also need a uniform lower-tail estimate for N. Let $\nu = pm \geq 3$ and $F = \{N \geq \nu/2\}$. Since $\mathbb{E}N^2 \leq \nu^2 + \nu$, the inequality

$$\nu = \mathbb{E}N \leq \nu/2 + \sqrt{\mathbb{E}N^2 \Pr(F)}$$

gives

$$\Pr(F) \geq \frac{\nu}{4(\nu+1)} \geq \frac{1}{6}, \qquad N \geq 2, \quad \log N \geq \tfrac{1}{2} \log \nu \quad \text{on } F. \tag{15}$$

To see the last two claims, if $3 \leq \nu \leq 4$, integrality gives $N \geq 2 \geq \sqrt{\nu}$ on F: if $\nu \geq 4$, then $N \geq \nu/2 \geq \sqrt{\nu}$.

The achieved gain. Under exact preferences, the subset rule selects its largest gold reward. The subset was drawn independently of rewards, so

$$\Delta(\pi) = \sigma \mathbb{E} a_s. \tag{16}$$

For $K \geq \log 2$, the inequality $\lfloor e^K \rfloor \geq e^{K/2}$, together with equation 11 and equation 15, implies, on F,

$$\log s \geq \tfrac{1}{2} \min\{K, \log(pm)\} \geq \tfrac{1}{4} h.$$

Since $a_s \geq 0$ also outside F, we obtain

$$\Delta(\pi) \geq \sigma \Pr(F) c_0 \sqrt{h/4} \geq \frac{c_0}{12} \sigma \sqrt{h}.$$

For $K < \log 2$, the sampled pair has independent Gaussian rewards and $\mathbb{E}|R_i - R_j| = 2\sigma/\sqrt{\pi}$. Biasing the uniform choice toward the better response by $\theta = \sqrt{K}$ therefore gives

$$\Delta(\pi) = \frac{\sigma\sqrt{K}}{\sqrt{\pi}} \Pr(N \geq 2) \geq \frac{\sigma\sqrt{K}}{6\sqrt{\pi}}. \tag{17}$$

Here $h = K$, since $q \geq 4$. Both ranges are thus covered by the universal constant

$$c_{\mathrm{ex}} := \min\left\{ \frac{c_0}{12}, \frac{1}{6\sqrt{\pi}} \right\} > 0. \tag{18}$$

At $K = 0$, the pair rule is the base policy and the claimed lower bound is zero. Applying equation 10 completes the proof of Theorem 2.

A.3 Proof of Theorem 3

The upper bound follows from Appendix A.1. For the lower bound, remember the set S, base policy, and notation from Appendix A.2, with $q \geq 4$. The subset and pair rules remain feasible for any ordering of S. We proceed by bounding the reward lost when this ordering is obtained from a random preference dataset.

Estimating the ranking. For $i \in S$, let w_i be its fraction of wins and let

$$u_i = \mathbb{E}[w_i \mid R] = \frac{1}{m-1} \sum_{j \in S \setminus \{i\}} \ell\left(\frac{R_i - R_j}{\beta}\right), \qquad \ell(x) = \frac{1}{1+e^{-x}}.$$

Order responses by w_i, breaking ties by index. Set

$$b_m = 2\sigma\sqrt{\log m}, \qquad t_m = \sqrt{\frac{2\log m}{m-1}},$$

and define

$$G_R = \left\{\max_{i \in S} |R_i - \mu| \leq b_m\right\}, \qquad G_C = \left\{\max_{i \in S} |w_i - u_i| \leq t_m\right\}.$$

The Gaussian tail bound and a union bound give $\Pr(G_R^c) \leq 2/m$. Conditional on R, each w_i averages $m-1$ independent Bernoulli variables. Hoeffding's inequality and another union bound therefore give

$$\Pr(G_C^c \mid R) \leq 2m \exp\{-2(m-1)t_m^2\} = \frac{2}{m^3}.$$

The scores of different responses need not be independent for this bound.

On G_R, every argument of ℓ in these scores lies in $[-2b_m/\beta, 2b_m/\beta]$. Since $\ell'(x) \geq e^{-|x|}/4$, if $R_i \geq R_j$ then

$$\begin{aligned}(m-1)(u_i - u_j) &= \ell\left(\frac{R_i - R_j}{\beta}\right) - \ell\left(\frac{R_j - R_i}{\beta}\right) \\ &\quad + \sum_{k \in S \setminus \{i,j\}} \left[\ell\left(\frac{R_i - R_k}{\beta}\right) - \ell\left(\frac{R_j - R_k}{\beta}\right)\right] \\ &\geq \frac{m e^{-2b_m/\beta}}{4\beta}(R_i - R_j). \end{aligned} \tag{19}$$

The mutual comparison contributes two terms to this lower bound, each of the other $m-2$ opponents contributes one. If $w_j \geq w_i$ despite $R_i \geq R_j$, then $u_i - u_j \leq 2t_m$ on G_C. Consequently, on $G_R \cap G_C$, every such inverted pair satisfies

$$R_i - R_j \leq \delta_m, \qquad \delta_m = 8\beta e^{2b_m/\beta} t_m. \tag{20}$$

This holds simultaneously for every pair. In particular, in any nonempty subset of S, the response with the largest score has reward within δ_m of the largest true reward.

To control the loss off this event, let $D_m = \max_{i \in S} R_i - \min_{i \in S} R_i$ and $U_m = \max_{i \in S} |R_i - \mu|/\sigma$. A Gaussian union bound gives

$$\mathbb{E}U_m^2 \leq \int_0^\infty \min\{1, 2me^{-v/2}\}\, dv = 2\log(2m) + 2,$$

so $\mathbb{E}D_m^2 \leq 8\sigma^2(\log(2m)+1)$. Since $\Pr((G_R \cap G_C)^c) \leq 4/m$, Cauchy–Schwarz shows that the expected loss from maximizing the score rather than the reward in any subset is at most

$$\eta_m = 8\beta e^{(4\sigma/\beta)\sqrt{\log m}} \sqrt{\frac{2\log m}{m-1}} + 4\sqrt{2}\sigma\sqrt{\frac{\log(2m)+1}{m}}. \tag{21}$$

The bound is uniform over the subset, and hence applies to the random subsets used by our policy. It also applies to a random pair. Empty admissible sets within S use the common fallback and incur no loss.

The resulting reward gain. For $K \geq \log 2$, draw a uniform subset of B of size $s = \min\{N, \lfloor e^K \rfloor\}$ and select its highest-scoring member. The subset is sampled independently of rewards and comparison data conditional on A. Equation 16 and the uniform loss bound imply

$$\Delta(\pi) \geq \sigma \mathbb{E} a_s - \eta_m, \tag{22}$$

where $s = 0$ on $N = 0$ and $a_0 = 0$. For $0 \leq K < \log 2$, favor the highest-scoring member of a pair sampled uniformly from B with bias $\sqrt{K}$. This rule is the mixture of selecting that member with probability $\sqrt{K}$ and selecting uniformly from the pair otherwise. Its loss relative to the exact-ranking rule is therefore at most $\sqrt{K}\,\eta_m$. By equation 17,

$$\Delta(\pi) \geq \sqrt{K}\left(\frac{\sigma}{\sqrt{\pi}} \Pr(N \geq 2) - \eta_m\right). \tag{23}$$

The base policy is used when $N \leq 1$. The factor $\sqrt{K}$ on the error makes this estimate uniform as K decreases to zero.

A bound for every dataset size. Set $r = \beta/\sigma$. Dividing equation 21 by σ gives

$$\frac{\eta_m}{\sigma} = 8re^{(4/r)\sqrt{\log m}}\sqrt{\frac{2\log m}{m-1}} + 4\sqrt{2}\sqrt{\frac{\log(2m)+1}{m}} \longrightarrow 0 \tag{24}$$

for every fixed $r > 0$: the exponential factor is $m^{o(1)}$. Recall that $c_{\mathrm{ex}} \leq 1/(6\sqrt{\pi})$ by equation 18. Choose an integer $m_0(r) \geq 4$ such that, for all $m \geq m_0(r)$,

$$\frac{\eta_m}{\sigma} \leq \min\left\{\frac{c_{\mathrm{ex}}\sqrt{\log 2}}{2}, \frac{1}{12\sqrt{\pi}}\right\}.$$

This threshold depends only on r. When $K \geq \log 2$, we have $h \geq \log 2$, and the deterministic subset bound together with equation 22 gives $\Delta(\pi) \geq (c_{\mathrm{ex}}/2)\sigma\sqrt{h}$. When $K < \log 2$, we have $h = K$; using $\Pr(N \geq 2) \geq 1/6$ in equation 23 gives the same conclusion.

To cover the remaining finite range $m < m_0(r)$, we sample a pair uniformly from B and use its observed preference, following the base policy if $N < 2$. If $K \geq \log 2$, select the preferred member, otherwise we favor it with bias $\sqrt{K}$. The pair calculation in the deterministic proof gives divergence at most K for both rules, regardless of the label. If $D \sim \mathcal{N}(0, 2)$, define

$$\gamma(r) = \frac{1}{2}\mathbb{E}\left[D \tanh\left(\frac{D}{2r}\right)\right] > 0.$$

Conditioning on the two rewards shows that following their observed label has expected gain $\sigma\gamma(r)$. Favoring it with bias $\sqrt{K}$ multiplies this gain by $\sqrt{K}$. The expectation defining $\gamma(r)$ is finite, and its integrand is positive almost surely. Therefore

$$\Delta(\pi) \geq \frac{\sigma\gamma(r)}{6}\min\{\sqrt{K}, 1\}.$$

In this range $q \leq \sqrt{2M} < m + 1 \leq m_0(r) + 1$, so

$$\min\{\sqrt{K}, 1\} \geq \frac{\sqrt{h}}{\sqrt{\max\{1, \log(m_0(r)+1)\}}}.$$

Thus a positive constant valid at every finite budget is

$$c_{\mathrm{noisy}}(r) = \min\left\{\frac{c_{\mathrm{ex}}}{2}, \frac{\gamma(r)}{6\sqrt{\max\{1, \log(m_0(r)+1)\}}}\right\}.$$

All the constructed policies use the same reward-blind base and meet the divergence constraint. Equation 10 transfers their gain to every constrained proxy optimizer, proving the theorem.

Table 2: **Training and evaluation prompt topics.** Counts and percentages within each set, using original Arena annotations. Tags may overlap. The dataset has no finer topic annotations for prompts marked *Topic unspecified*.

Topic	Training ($n = 1,000$)	Evaluation ($n = 128$)
Coding	325 (32.5%)	43 (33.6%)
Mathematics	81 (8.1%)	14 (10.9%)
Creative writing	68 (6.8%)	8 (6.3%)
Topic unspecified	558 (55.8%)	67 (52.3%)

A.4 Proof of Corollary 4

We compare every pair in S and use the same subset rule as above. Along the sequences in the corollary, $q = p\sqrt{2M} \to \infty$ and $K \to \infty$. Since $q - 1 \le pm \le q$, we have $pm/q \to 1$ and $m \to \infty$. Write

$$h_m := \min\{K, \log(pm)\} = h + o(1), \qquad h \to \infty.$$

A binomial Chernoff bound gives $\Pr(N \ge pm/2) \ge 1 - e^{-pm/8}$. On this event, for all sufficiently large budgets,

$$s = \min\{N, \lfloor e^K \rfloor\} \ge \tfrac{1}{2} \min\{pm, e^K\} = \tfrac{1}{2} e^{h_m}.$$

Monotonicity and nonnegativity of a_j, followed by equation 14, give

$$\mathbb{E}a_s \ge (1 - e^{-pm/8}) a_{\lfloor e^{h_m}/2 \rfloor} \ge (1 - o(1))\sqrt{2h}. \tag{25}$$

The exact-preference construction has gain $\sigma \mathbb{E}a_s$. For noisy preferences, the construction loses at most η_m, as established in equation 22. For fixed $\sigma, \beta > 0$, $\eta_m = o(1)$, so this loss is negligible relative to $\sigma\sqrt{h}$. Thus either preference model admits a feasible policy with gain

$$(1 - o(1))\sigma \sqrt{2 \min\{K, \log(p\sqrt{2M})\}}.$$

Thus the posterior-mean proxy transfers this bound to every optimizer by equation 10.

B Additional Experimental Details

B.1 Data construction and topic composition

We use English first-turn prompts from LMArena's `arena-human-preference-140k`. The final training set contains 1,000 prompts, each contributing 300 disjoint response pairs and hence 600 retained responses. Responses are sampled from Qwen3-4B-Instruct-2507 with temperature 1, top-$p = 1$, no top-k restriction, and at most 1,024 new tokens. Pair construction excludes identical responses and repeated text pairs, additional responses are sampled when necessary to obtain 300 eligible pairs.

Table 2 gives the prompt-topic distribution.

B.2 Conditional gold-reward distributions

The theoretical model in Section 3.1 assumes Gaussian rewards. We examine this idealization using the empirical distribution of gold scores within individual prompts. Each example uses all 8,192 responses in its test-prompt pool, scored by the fixed gold reward model before best-of-N selection.

We choose four prompts independently of their reward values. The examples concern AutoHotkey scrolling, a modal-logic agent, motivational story topics, and railway recruitment exam questions.

For a prompt with scores $r_1, \ldots, r_n$, let $\bar{r} = n^{-1} \sum_i r_i$ and $s^2 = n^{-1} \sum_i (r_i - \bar{r})^2$. The normal Q–Q plot compares the ordered standardized scores $z_{(i)} = (r_{(i)} - \bar{r})/s$ with

$$q_i = \Phi^{-1}\left(\frac{i - 1/2}{n}\right), \qquad i = 1, \ldots, n, \tag{26}$$

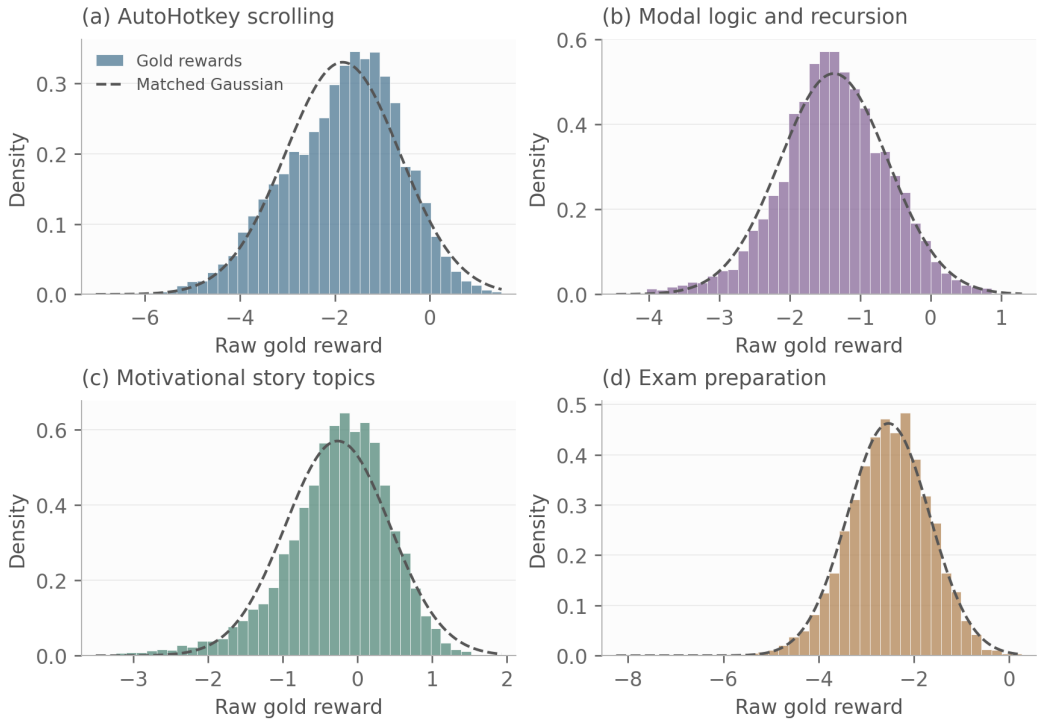


Figure 3: **Gold-reward distributions for four test prompts.** Each histogram contains all 8,192 raw gold scores for one prompt. The overlaid Gaussian has the same empirical mean and variance. Panels use the same prompt order as Figure 4; prompts are selected by topic and question ID without inspecting rewards.

where Φ is the standard normal cumulative distribution function. Gaussian scores would approximately follow the line $z = q$.

Figures 3 and 4 show that the conditional gold-reward distributions need not be Gaussian. All four examples have negative empirical skewness, ranging from -0.75 to -0.22. The creative-writing example has the strongest asymmetry, with skewness -0.75 and excess kurtosis 1.19. Departures from the reference line in the Q–Q plots reveal differences in distributional shape even after matching each prompt's mean and variance. Thus, the empirical scaling results arise in a setting whose reward distributions depart from the Gaussian idealization.

B.3 Hyperparameters

We found reward-model training to work over a broad range of learning rates in preliminary sweeps, from 2×10^{-6} to 4×10^{-5}, consistent with the findings of Ouyang et al. (2022, Appendix C.2). We then compared 2×10^{-6} and 10^{-5} in 75,000-comparison pilots, selecting by final logistic loss on 512 validation comparisons from 64 held-out prompts distinct from training and test sets. All validations selected 10^{-5}, which we use for the scaling experiments; Coste et al. (2024) use the same reward-model learning rate. Table 3 gives the complete shared settings.

C Additional Experiments

Figure 5 collects the square-root-logarithmic fits for all ten settings in Table 1.

Figure 6 shows the best-of-N reward curves for four comparison budgets, using 0.6B proxies trained on noiseless and noisy preferences.



Figure 4: **Normal Q–Q plots for the same four prompts.** Ordered gold scores, standardized within each prompt, are plotted against standard normal quantiles at probabilities $(i - 1/2)/8{,}192$. The diagonal is the Gaussian reference. Systematic departures, particularly in the tails, illustrate that empirical gold rewards need not follow a Gaussian distribution.

Table 3: **Training hyperparameters.**

Hyperparameter	Value
Trainable parameters	Full backbone and scalar reward head
Objective	Pairwise logistic loss
Optimizer	AdamW
Learning rate / schedule	10^{-5} / constant, no warmup
Adam (β_1, β_2) / ϵ	$(0.9, 0.999)$ / 10^{-8}
Weight decay	10^{-3} on non-normalization matrix weights
Gradient clipping	Global ℓ_2 norm of 1
Training epochs	1
Effective batch size	600 comparisons
Model precision	BF16 weights; FP32 gradient accumulation, master weights, and Adam states
Input format	Qwen chat template; thinking disabled
Maximum sequence length	4,096 tokens; no truncation, right padding
Attention / attention dropout	FlashAttention 2 / 0
Reward head	Bias-free linear head on the final nonpadding token
Head initialization	$\mathcal{N}(0, 0.02^2)$

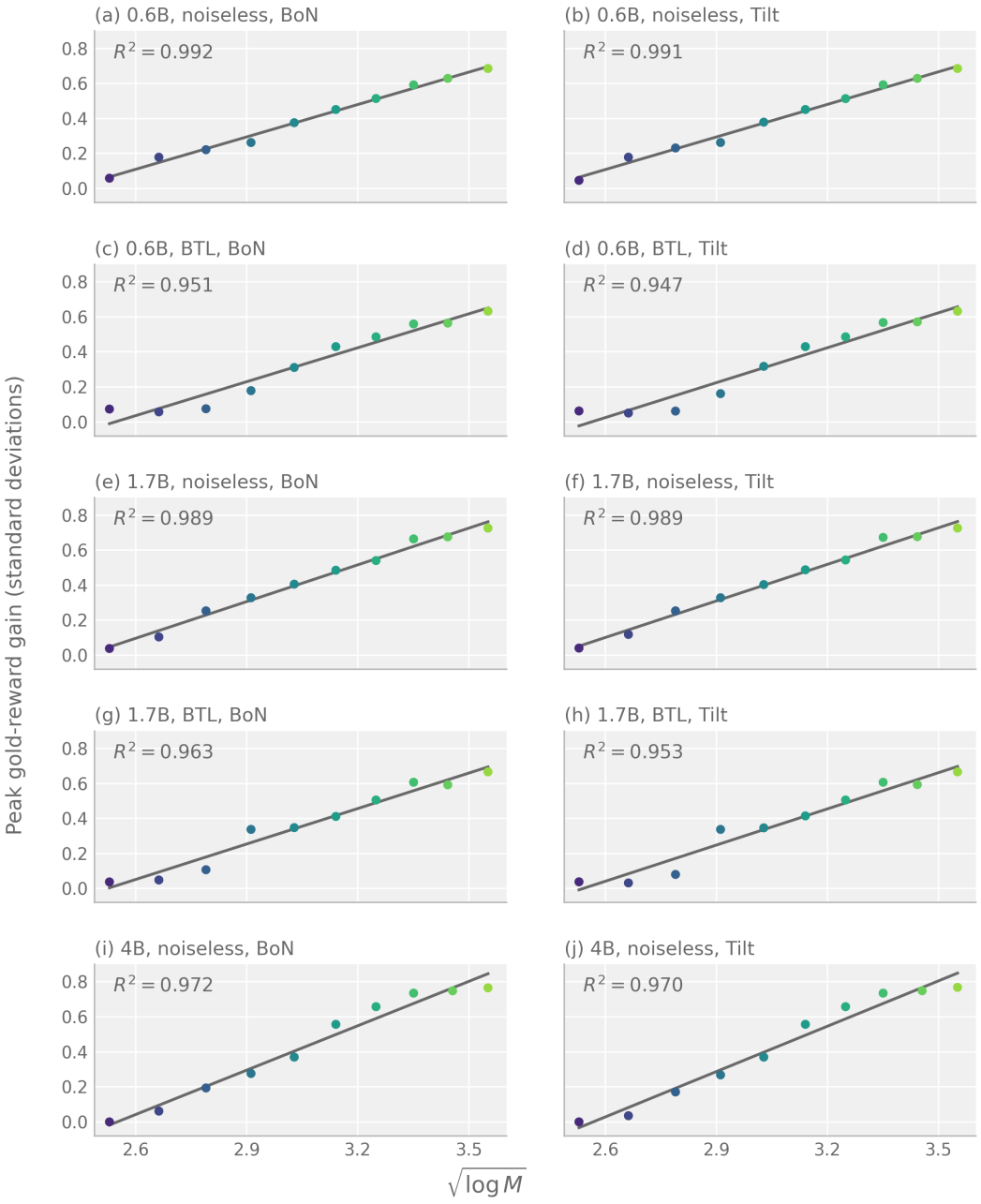


Figure 5: **Square-root-logarithmic scaling across all settings in Table 1.** Columns show best-of-N (BoN) and exponential tilting (Tilt); rows vary proxy size and preference noise. Each point is the peak mean gold-reward gain over the evaluated optimization grid. Lines fit $a + b\sqrt{\log M}$ to ten models between $M = 600$ and 300,000.

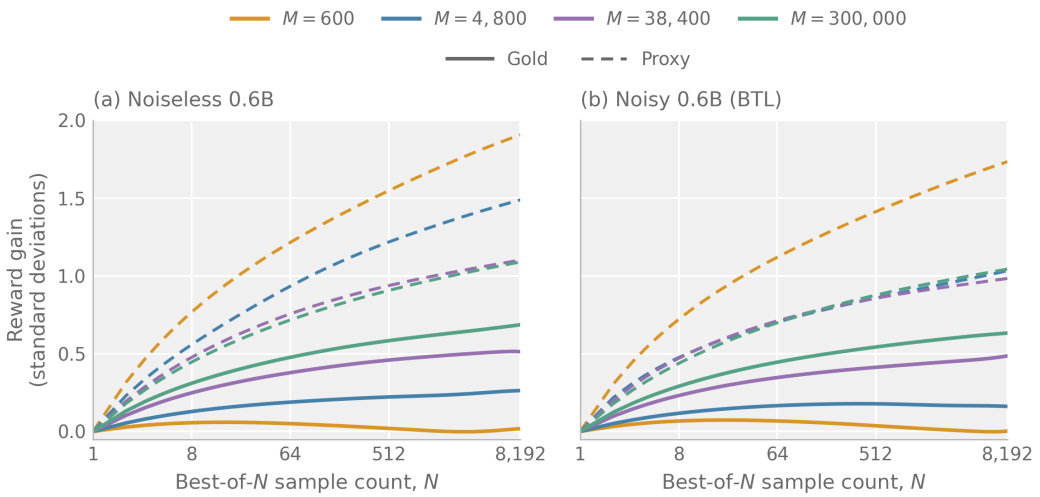


Figure 6: **Best-of-N optimization with noiseless and noisy preferences.** Each panel shows four 0.6B checkpoints trained on $M = 600$, 4,800, 38,400, and 300,000 comparisons. The N-axis is logarithmic. Solid and dashed lines show mean gold- and proxy-reward gains over the same evaluation prompts, using the unbiased subset-average estimator on 8,192 responses per prompt. Gains are centered at $N = 1$. Gold rewards use the fixed reference standard deviation, and each proxy uses its own standard deviation over the full evaluation pool.