Bayesian Preference Learning for Test-Time Steerable Reward Models

Jiwoo Hong[*]
LinkedIn Corporation

Shao Tang[*]
Nubank

Zhipeng Wang
Google LLC

Abstract

Reward models are central to aligning language models with human preferences via reinforcement learning (RL). As RL is increasingly applied to settings such as verifiable rewards and multi-objective alignment, RMs are expected to encode more complex and multifaceted preference distributions. However, classifier RMs remain static once trained, limiting their adaptability at test time. We propose **Variational In-Context Reward Modeling (ICRM)**, a novel Bayesian reward modeling objective that enables *test-time steerability* via in-context preference demonstrations. ICRM casts reward modeling as amortized variational inference over a latent preference probability under the Bradley-Terry model using a conjugate Beta prior. We show that ICRM adapts to unseen preference distributions at test time for both single and multi-objective settings. With more demonstrations, ICRM improves RM-Bench accuracy from 60.5 to 70.8, achieves lower calibration error than a generative judge on moral dilemma preferences, and expands the attainable Pareto frontier under conflicting preferences. We further study the practical applicability of ICRM for RL training, showing that it can effectively encode verifiable rewards by outperforming a conventional RM in math reasoning. Finally, we provide theoretical guarantees that the variational objective admits a global interior optimum with finite confidence, and we analyze how KL regularization mitigates reward over-optimization.

1 Introduction

Reward models (RMs) serve as essential proxies for human preferences in language model post-training, including reinforcement learning with human feedback (RLHF) [Ziegler et al., 2020, Ouyang et al., 2022, Stiennon et al., 2020]. Specifically, triplets comprising a prompt, a preferred response, and a dispreferred response are used to parameterize the preference distribution under the Bradley–Terry (BT) model [Bradley and Terry, 1952]. Neural classifiers, *i.e.,* classifier RMs, act as estimators of the BT strength parameter, with theoretical guarantees that, given sufficient preference data, the learned RM can approximate the "true" human preference distribution [Bong and Rinaldo, 2022, Rafailov et al., 2023]. This formulation enables the learned RM to act as a standalone proxy for a single concatenation of prompt and response, which is practically useful for RLHF training.

However, classifier RMs face two data-driven limitations: (1) they are *static* once trained on a given dataset, and (2) they are prone to over-optimization [Gao et al., 2023, Hong et al., 2025]. While LLM-as-a-Judge [Kim et al., 2024] offers flexible evaluation criteria with strong performance [Lambert et al., 2025, Malik et al., 2025, Liu et al., 2025b], these gains often rely on proprietary models [Comanici et al., 2025, OpenAI et al., 2024], implying substantial compute and data costs. Hence, it is desirable to design an efficient classifier RM that remains adaptable to unseen data while avoiding over-optimization by being test-time steerable.

[*]Equal Contribution

RLxF: Reinforcement Learning from World Feedback.

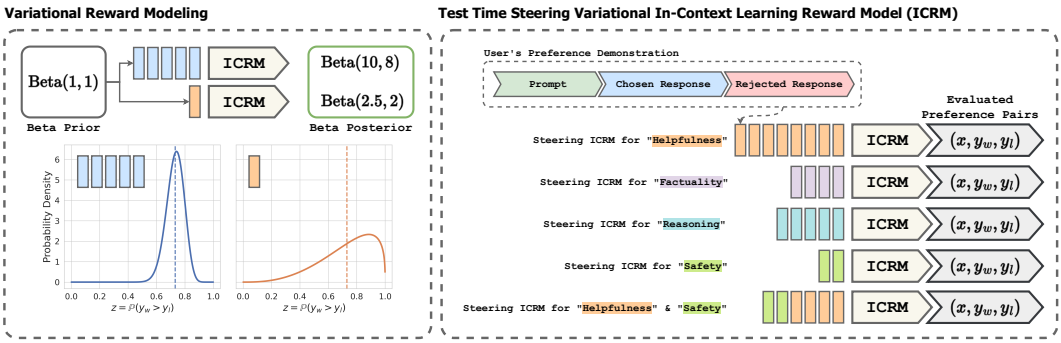


Figure 1: Variational in-context reward modeling (**ICRM**) with Beta prior for the Bradley-Terry (BT) model. ICRM directly encodes the mean and sharpness of the Beta posterior given in-context preference demonstrations, allowing **multi-objective test-time steerability** for *any* unseen preferences.

In this paper, we introduce a **variational in-context reward modeling (ICRM)** framework grounded in a Bayesian view of preferences. Our method approximates the true preference distribution with a Beta posterior conditioned on in-context preference demonstrations. In detail, placing a Beta prior on the BT model yields a closed-form training loss via variational inference. This variational loss enables ICRM to learn preferences *in-context* with few-shot demonstrations, allowing **test-time steerability of a classifier RM** that can dynamically adapt to one or more mixtures of arbitrary preferences, as in Figure 1. Furthermore, we prove that a KL penalty to the Beta prior tempers the learned preference mean and yields a global interior optimum. Our main contributions are summarized below:

1. **Principled variational preference learning**: We propose a reward modeling objective that enables RMs to infer test-time preferences from in-context preference samples. We prove that regularizing the Beta posterior with a uniform Beta prior guarantees a global interior optimum, tempering excessive maximization of the preference mean on training data.
2. **Single and multi-objective test-time steerability**: We show that the variational design yields both directional and calibrated steerability, and extends to multi-objective preference trade-offs. With an increasing number of demonstrations, ICRM improves RM-Bench accuracy by over 10% point and increases hypervolume by 4% on conflicting RM-Bench Safety subsets.
3. **Parameterizing verifiable rewards with ICRM**: We apply ICRM to reinforcement learning with verifiable rewards (RLVR), where it outperforms both the verifier and the Bradley-Terry reward model on math reasoning, achieving over 2.5% point accuracy compared to the gold verifier.

2 Background

Preliminaries. A classifier reward model (RM), $r_\theta(x, y)$, is a function parameterized by θ that outputs a score indicating the quality of a prompt-response pair (x, y) [Ziegler et al., 2020]:

$$r_\theta(x, y) = W_p^\top h_\theta(x, y) \in \mathbb{R}, \tag{1}$$

where $W_p \in \mathbb{R}^{d_{\text{model}} \times 1}$ is a projection head initialized by $\mathcal{N}(0, (d_{\text{model}} + 1)^{-1})$ [Stiennon et al., 2020, Huang et al., 2024, Hong et al., 2025] and $h_\theta(x, y) \in \mathbb{R}^{d_{\text{model}} \times 1}$ is the last hidden state from the backbone language model. These models are typically trained on a dataset of human preferences, $\mathcal{D} = \{(x_i, y_{i,w}, y_{i,l})\}_{i=1}^N$, where $y_{i,w}$ is the preferred ("chosen") response and $y_{i,l}$ is the dispreferred ("rejected") response for a given prompt x_i. The training objective maximizes the log-likelihood of the preferences according to the Bradley-Terry (BT) model [Bradley and Terry, 1952],

$$P(y_w \succ y_l \mid x) = \sigma(r_\theta(x, y_w) - r_\theta(x, y_l)) = \frac{\exp(r_\theta(x, y_w))}{\exp(r_\theta(x, y_w)) + \exp(r_\theta(x, y_l))}, \tag{2}$$

which posits that the probability of y_w being preferred over y_l is given by a logistic function of the difference in their reward scores. The final loss function $\mathcal{L}_{\text{BT}}(\theta)$ is defined as:

$$\mathcal{L}_{\text{BT}}(\theta) = -\left(\log \sigma(r_\theta(x, y_w) - r_\theta(x, y_l))\right). \tag{3}$$

Once the preference distribution shown in the training set is encoded into θ via fine-tuning, it cannot be adaptively updated at test time without additional training, limiting the flexibility of RMs.

Estimating the true preference distribution. Prior work in offline preference learning supports that, with sufficient pairwise comparisons, fitted models recover underlying preferences [Rafailov et al., 2023, Hejna et al., 2024]. In the classical BT setting, the maximum-likelihood estimator (MLE) exists and enjoys consistency and asymptotic normality. For any context x and pair (y_w, y_l), if $\hat{P}$ is the probability estimated by the MLE and P^* the true probability, then $\hat{P}(y_w \succ y_l \mid x) \xrightarrow{p} P^*(y_w \succ y_l \mid x)$. Thus, with sufficient data, a learned BT model converges to the *true* preference distribution. We study the parameterization of the BT model and its applications in Appendix A.

Bayesian treatment of the Bradley-Terry model. Bayesian treatment of the BT model necessitates the selection of a suitable prior distribution for parameters [Chen and Smith, 1984, Whelan, 2017, Wainer, 2023, Fageot et al., 2024]. The general form of the model with $N > 2$ contenders is parameterized by a vector of N strength scores $\beta = (\beta_1, \ldots, \beta_N) \in \mathbb{R}^N$. Typically, the Bayesian formulation in those cases defines the prior distributions directly on each strength parameter: *e.g.,* Gaussian prior [Wainer, 2023] and a Dirichlet prior [Chen and Smith, 1984].

3 Variational In-Context Reward Modeling

We present a novel Bayesian reward modeling objective by framing in-context reward modeling as a problem of amortized variational inference. The central idea is to approximate the true preference distribution with a Beta posterior conditioned on in-context preference demonstrations.

3.1 Problem Setup

Prior. We introduce a latent random variable z represents the probability of y_w being preferred over y_l given prompt x and demonstrations $\mathcal{C}$, *i.e.,* $z \coloneqq P(y_w \succ y_l \mid x, \mathcal{C}) \in [0, 1]$. This captures the *preference standard* specific to the pair (y_w, y_l) under context $\mathcal{C}$ and prompt x. We assume there exists a true but intractable *context-dependent* prior, $p(z \mid x, y_w, y_l, \mathcal{C})$, that reflects implicit preference functions learned in context. Conditioned on z, the likelihood of the observed outcome $o \in \{0, 1\}$, $p(o \mid z, x, y_w, y_l, \mathcal{C})$, belongs to the Bernoulli family.

Posterior. By Bayes' rule, the *true posterior* over z after observing o, *i.e.,* our inferential target is:

$$p(z \mid o, \cdot) \propto p(o \mid z, \cdot) \times p(z \mid \cdot), \tag{4}$$

where $(x, y_w, y_l, \mathcal{C})$ is omitted. However, this is intractable as the context-dependent prior $p(z \mid x, y_w, y_l, \mathcal{C})$ lacks a simple analytical form due to the complex dynamics of in-context learning. Throughout, we focus on $o = \mathbb{1}_{y_w \succ y_l}$. Thus, we approximate the posterior distribution through $q_\theta(z \mid o = \mathbb{1}_{y_w \succ y_l}, x, y_w, y_l, \mathcal{C})$, which is denoted as $q_\theta(z \mid x, y_w, y_l, \mathcal{C})$ for notational brevity.

3.2 Reward Modeling as Variational Inference

We parameterize $q_\theta(z \mid x, y_w, y_l, \mathcal{C})$ with model θ, which directly maps the inputs to the parameters of an approximate posterior distribution, namely the variational in-context reward modeling (ICRM). We outline the choice of the prior distribution and propose the final learning objective for ICRM.

Beta prior for the Bradley-Terry model. Building on the discussion of the Bayesian treatment of the BT model, we propose using a Beta prior for the BT model in reward modeling. The setting for reinforcement learning from human feedback (RLHF) typically involves a *single* pairwise comparison (y_w, y_l) given the prompt x [Wang et al., 2024b, Liu et al., 2025a]. This specialization to $N = 2$ significantly reduces the problem's complexity, *i.e.,* likelihood of observing preference outcomes for this pair follows a Bernoulli distribution parameterized by z. For a Bernoulli likelihood, the conjugate prior for the parameter is the Beta distribution: $\mathrm{Beta}\,(\alpha_0, \beta_0)$, where (α_0, β_0) encodes our initial belief about the preference before observing any data.

Amortized variational approximation of posterior. Given the Beta prior, we approximate the posterior distribution $q_\theta(z \mid x, y_w, y_l, \mathcal{C})$ using a reward model with a two-dimensional projection head $W_p \in \mathbb{R}^{d_{\text{model}} \times 2}$, returning a *utility* score $u_\theta(x, y, \mathcal{C})$ and a *confidence* (*i.e.,* evidence) score $s_\theta(x, y, \mathcal{C})$, which are context dependent. For $(x, y_w, y_l, \mathcal{C})$, we have both scores, each for the chosen

and rejected responses y_w and y_l, shortened as u_w, u_l, s_w, and s_l. We reparameterize the Beta posterior $\mathrm{Beta}(\alpha_q, \beta_q)$ with $\alpha_q = \mu\tau$ and $\beta_q = (1-\mu)\tau$, where

$$\mu = \sigma(u_w - u_l) \text{ and } \tau = \mathrm{Softplus}(s_w) + \mathrm{Softplus}(s_l) + 1, \tag{5}$$

with $\mathrm{Softplus}(x) = \log(1 + \exp(x))$. Here $\mu \in (0, 1)$ is the posterior predictive probability and $\tau > 0$ controls concentration. The approximate posterior is $q_\theta(z \mid x, y_w, y_l, \mathcal{C}) = \mathrm{Beta}\big(z; \alpha_q, \beta_q\big)$, with $\alpha_q > 0$ and $\beta_q > 0$. This construction preserves the BT model as a special case: the posterior mean of q_θ recovers the BT preference probability, $\mathbb{E}_{q_\theta}[z] = \alpha_q/(\alpha_q + \beta_q) = \mu = \sigma(u_w - u_l)$, while the concentration τ reflects the amount of evidence.

Evidence lower bound for variational objective. Since the true posterior $p(z \mid x, y_w, y_l, \mathcal{C})$ is intractable as described in Section 3.1, we formulate the inference task as an optimization problem using variational inference to approximate the true posterior with the reward model r_θ. Inspired by Joo et al. [2020], we train the model θ by maximizing the Evidence Lower Bound (ELBO) for the observed preference $y_w \succ y_l$. The loss is the negative ELBO:

$$\mathcal{L}_{\mathrm{ELBO}}(\theta) = -\underbrace{\mathbb{E}_{q_\theta(z|x,y_w,y_l,\mathcal{C})}\left[\log z\right]}_{\text{Reconstruction Error}} + \lambda(N) \times \underbrace{\mathbb{D}_{\mathrm{KL}}\left(q_\theta\left(z \mid \cdot\right) \mid\mid p\left(z \mid \cdot\right)\right)}_{\text{Regularization Term}}, \tag{6}$$

where $\cdot$ in the regularization term omits $(x, y_w, y_l, \mathcal{C})$. The first term in (6) represents the *reconstruction error*, measuring how well the approximate posterior explains the observed outcome $y_w \succ y_l$. For a Beta distribution, this expectation has a known closed form solution involving the digamma function, $\psi(x) := d \log \Gamma(x)/dx$:

$$\mathbb{E}_{q_\theta(z|x,y_w,y_l,\mathcal{C})}\left[\log z\right] = \psi(\alpha_q) - \psi(\alpha_q + \beta_q) = \psi(\mu\tau) - \psi(\tau). \tag{7}$$

Minimizing this term increases μ toward 1, favoring y_w, analogous to the standard BT loss [Azar et al., 2024]. Meantime, τ controls how sharply the distribution concentrates around this preference.

The second term in (6) is the Kullback-Leibler (KL) divergence from the model's approximate posterior $q_\theta = \mathrm{Beta}(\mu\tau, (1-\mu)\tau)$ to the prior p. As the true prior $p(z \mid x, y_w, y_l, \mathcal{C})$ is intractable, we replace it with a fixed, uninformative prior $p(z) = \mathrm{Beta}(z; \alpha_0, \beta_0)$, *e.g.,* a uniform prior with $\alpha_0 = \beta_0 = 1$. And $\lambda(N)$ is a monotonically decreasing schedule that down-weights the KL term as the amount of contextual evidence N grows. This term regularizes the approximation, preventing the posterior from deviating excessively from the prior, especially when contextual evidence is minimal, *e.g.,* N is small. The KL divergence between two Beta distributions, $p = \mathrm{Beta}(\alpha_p, \beta_p)$ and $q = \mathrm{Beta}(\alpha_q, \beta_q)$, has a closed form solution [Loaiza-Ganem and Cunningham, 2019]:

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(q \mid\mid p) = \log \frac{\Gamma(\alpha_q + \beta_q)}{\Gamma(\alpha_q)\Gamma(\beta_q)} - \log \frac{\Gamma(\alpha_p + \beta_p)}{\Gamma(\alpha_p)\Gamma(\beta_p)} + (\alpha_q - \alpha_p)[\psi(\alpha_q) - \psi(\alpha_q + \beta_q)] \\ + (\beta_q - \beta_p)[\psi(\beta_q) - \psi(\alpha_q + \beta_q)],\end{aligned} \tag{8}$$

Finally, the dynamic hyperparameter $\lambda(N)$ controls this balance: when the context is minimal ($N = 1$), a large $\lambda(1)$ forces the posterior to remain close to the uninformative prior, *i.e.,* high uncertainty. As more examples are added to the context, $\lambda(N)$ decreases, allowing the reconstruction term to dominate and the model to form a more confident, data-driven posterior distribution. Combining these components, the fully-specified loss is defined as:

$$\mathcal{L}_{\text{ICRM}}(\mu, \tau; \alpha_0, \beta_0) = -\left(\psi(\mu\tau) - \psi(\tau)\right) + \lambda(N)\mathbb{D}_{\mathrm{KL}}\left(\mathrm{Beta}(\mu\tau, (1-\mu)\tau) \| \mathrm{Beta}(\alpha_0, \beta_0)\right), \tag{9}$$

where μ, τ are functions of θ and $\lambda(N) = \lambda \times N^{-1}$ with predefined λ. For notational convenience, we henceforth write $\mathcal{L}_{\text{ICRM}}(\mu, \tau)$.

Choice of uniform Beta prior for the divergence penalty. As in (9), the divergence penalty can be controlled with the predefined prior distribution $p = \mathrm{Beta}(\alpha_0, \beta_0)$. If we have explicitly collected annotations for the pair (x, y_w, y_l) for given few-shot examples $\mathcal{C}$, we may set unique (α_0, β_0). However, it is typically hard to collect such data. Thus, we assume $(\alpha_0, \beta_0) = (1, 1)$, implying the uniform distribution on preferring y_w over y_l without any information. Potentially, synthetic personas [Singh et al., 2025] or voting over multiple preference models [Yang et al., 2024c] can be used to generate such data to provide a more informative prior.

3.3 Analysis

One common failure mode of the Bradley-Terry (BT) reward model is *over-optimization* [Gao et al., 2023], in which the preference probabilities converge to 1 and fit into the local optima of the *true* human preference distribution [Azar et al., 2024, Hong et al., 2025]. The proposed KL-regularized variational objective directly addresses this issue, *i.e.,* it precludes boundary minima—ensuring an interior optimum—and, via the same KL term, imposes a quantitative edge-behavior barrier that moderates the excessive growth of the score margin at high preference probabilities.

Lemma 3.1 (Edge behavior at finite confidence)**.** *Let $P_\theta(y_w \succ y_l \mid x)$ denote the ICRM preference with $\mu = \sigma(\Delta u_\theta) = \sigma(u_\theta(x, y_w) - u_\theta(x, y_l))$ and $\varepsilon := 1 - \mu$. For $\tau \in (0, \infty)$, as $\varepsilon \to 0^+$,*

$$\nabla_\theta \mathcal{L}_{\text{ICRM}} \to \underbrace{\left(\frac{\lambda\beta_0}{\varepsilon\tau}\right)}_{\text{Utility Coefficient}} \nabla_\theta \Delta u_\theta + \underbrace{\left(-\frac{\lambda\beta_0}{\varepsilon\tau^2}\right)}_{\text{Confidence Coefficient}} \nabla_\theta \tau.$$

Here, Δu_θ is regularized by τ. As training increases μ, $\lambda\beta_0/(\tau\varepsilon)$ in the utility coefficient increases for any finite τ, thereby penalizing further growth of the utility and preventing uncontrolled maximization of μ when taking a gradient descent step (see Appendices F and G for proof). Since the Lemma 3.1 is for finite τ, we next prove that the global minimizer indeed has $0 < \tau^\star < \infty$.

Theorem 3.2. *Assume $\lambda > 0$ and $\alpha_0, \beta_0 > 0$. Every global minimizer $(\mu^\star, \tau^\star)$ of $\mathcal{L}_{ICRM}(\mu, \tau; \alpha_0, \beta_0)$ defined in* (9) *satisfies*

$$0 < \mu^\star < 1 \qquad \textit{and} \qquad 0 < \tau^\star < \infty.$$

Consequently, this provides a theoretically guaranteed prevention of reward model over-optimization via preference mean tempering. See Appendix H for proof.

4 Experiments

Given a *single* trained ICRM, we validate the test-time steerability from two aspects: (1) directional steerability and (2) calibrated steerability. Then, we conduct a case study on a multi-objective setting.

1. **Calibrated Steerability (single-objective)**: Can ICRM steer its posterior predictive mean μ in proportion to the preference mixture specified by in-context demonstrations $\mathcal{C}$?
2. **Directional steerability (single-objective)**: Does the posterior predictive mean $\mathbb{E}_{q_\theta}[z] = \mu$ adapt to the implicit preference distribution induced by in-context demonstrations $\mathcal{C}$?
3. **Multi-Objective Steerability**: Can the posterior predictive mean μ encode competing preferences with respect to the in-conttext demonstrations $\mathcal{C}$?

4.1 Training Configurations

We use Qwen3-4B [Yang et al., 2025] and Llama-3.2-3B [Dubey et al., 2024]. We use the pre-trained base models to prevent biased prior preference distributions from post-training. The projection head $W_p \in \mathbb{R}^{d_{\text{model}} \times 2}$ is initialized with $\mathcal{N}(0, (d_{\text{model}} + 1)^{-1})$ [Huang et al., 2024]. We use Skywork-Preferences-v0.2 [Liu et al., 2024], assuming each dataset reflects a consistent implicit preference distribution. For each training instance, we construct in-context demonstrations $\mathcal{C} = \{(x, y_w, y_l)\}_{j=1}^{N}$ with $N \in \{1, 2, 4, 8, 16\}$, sampled within train data. We use a prompt template without any arbitrary instructions in Appendix B to minimize template bias. We sweep different $\lambda \in \{0.1, 0.5, 1.0\}$ and report the training dynamics in the Appendix C, along with the detailed training details.

4.2 Evaluation Configurations

Calibrated steerability. We test whether ICRM can encode the *degree* of a preference shift specified by the composition of $\mathcal{C}$ on challenging human-generated moral dilemma preference pairs from princi/pal [Wang, 2025]. Given $\mathcal{C}$ drawn from SafeRLHF, we vary the preference mixture by flipping an r fraction of demonstrations given N, where $r \in \mathcal{R} = \{0, 0.1, \ldots, 1.0\}$. Let $\hat{p}_{\text{flip}}(r)$ denote the accuracy as a realized rate at which the model follows the flipped preference on held-out pairs. We measure calibrated steerability with:

$$\text{ECE} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} |\hat{p}_{\text{flip}}(r) - r|, \quad \text{RMSE} = \sqrt{\frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} (\hat{p}_{\text{flip}}(r) - r)^2}. \tag{10}$$

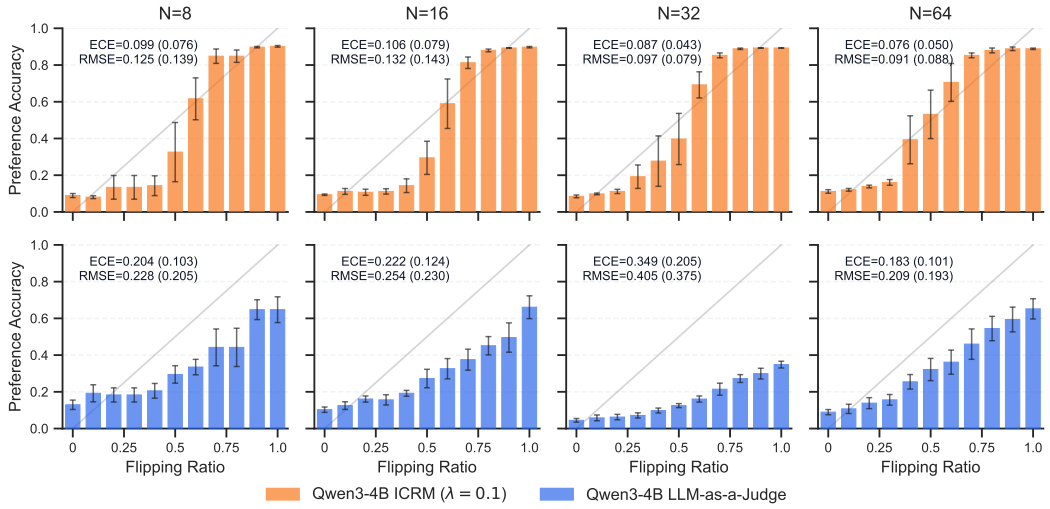


Figure 2: **Calibrated steerability.** Calibration analysis with ECE (↓) and RMSE (↓) between the flipping ratio of demonstrations and the preference accuracy on moral dilemma preference pairs.

Directional steerability. We use SafeRLHF [Ji et al., 2023], HHH Alignment [Srivastava et al., 2023], and RM-Bench [Liu et al., 2025b]. For the first two benchmarks and four domains in RM-Bench, we randomly sample $N \in \{1, 2, 4, 8, 16, 32, 64\}$ pairs to construct $\mathcal{C}$ within each domain. We then compare the scores for (x', y'_w) and (x', y'_l) conditioned on $\mathcal{C}$ for preference accuracy, where (x', y'_w, y'_l) is drawn from the held-out data. We report the mean and standard deviation of four runs.

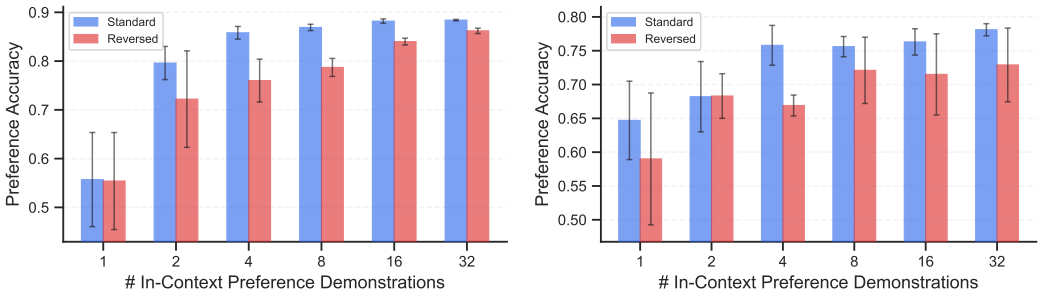
Multi-objective steerability. We use the *Safety-Should-Respond* and *Safety-Should-Refuse* subsets of the RM-Bench Safety domain, corresponding to benign and adversarial prompts. We construct $\mathcal{C}$ by mixing demonstrations from the two subsets with ratio $r \in [0, 1]$, where r denotes the proportion of helpfulness examples. For each N, we trace the Pareto frontier and report hypervolume [Zitzler and Thiele, 1999, HV] relative to $(0.0, 0.0)$. HV is reported only for methods that induce a frontier, since static RMs correspond to single operating points.

Baselines. We add five reward models (RMs) with different modeling objectives, which were trained with the open-source datasets. First, we compare with three classifier RMs, Bradley-Terry [Liu et al., 2024], ArmoRM [Wang et al., 2024a], and GRM [Yang et al., 2024d]. Additionally, we test the two generative RM approaches: plain generative use of post-trained Qwen3-4B ("LLM-as-a-Judge" in Table 1) [Zheng et al., 2023a] and GRAM-r^2 [Wang et al., 2026]. We use the default prompt template of RM-Bench for LLM-as-a-Judge.

5 Results

5.1 Calibrated Steerability (Single-Objective)

Figure 2 evaluates whether ICRM can follow the degree of preference shift specified by the composition of $\mathcal{C}$. As the flipping ratio r increases, ICRM consistently increases the rate of selecting the flipped preference on moral dilemma pairs, showing a monotonic response to the preference mixture. The resulting curves are closer to the ideal diagonal calibration curve than the Qwen3-4B LLM-as-a-Judge baseline across all tested context sizes. The calibration errors show the same trend. Across $N \in \{8, 16, 32, 64\}$, ICRM obtains lower ECE and RMSE than the baseline, with ECE decreasing from 0.099 at $N = 8$ to 0.076 at $N = 64$. These results indicate that ICRM can encode graded preference mixtures at test time, rather than only steering toward a single preference direction. Based on the extended calibrated steerability analysis for different λ and base models in Appendix D, we select Qwen3-4B ICRM with $\lambda = 0.1$ as the best model by having the smallest ECE and RMSE.



(a) HHH-Alignment [Srivastava et al., 2023]

(b) SafeRLHF [Ji et al., 2023]

Figure 3: **Directional Steerability.** The accuracy mean and standard deviation of ICRM with increasing number of in-context demonstrations from held-out set on HHH-Alignment and SafeRLHF.

Table 1: RM-Bench evaluation results with the standard deviation of four random seeds for ICRM.

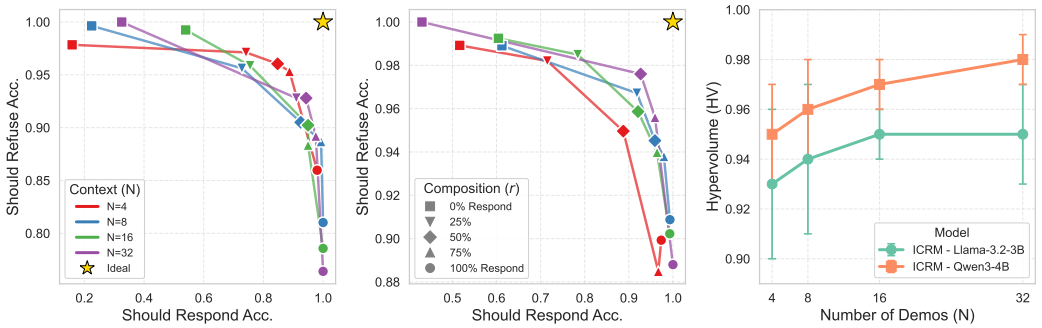
	RM-Bench				
	Chat	**Safety**	**Code**	**Math**	**Avg.**
Generative					
LLM-as-a-Judge	69.2	90.9	54.3	**67.6**	70.3
GRAM-r^2	61.1	92.9	54.7	61.6	67.6
Classifier					
BTRM	69.3	**96.0**	53.2	62.1	70.2
ArmoRM	67.8	92.4	53.1	57.5	67.7
GRM	62.7	90.0	**57.8**	62.5	68.2
ICRM (*Ours*)					
$N = 1$	$48.2_{12.6}$	$85.6_{15.4}$	$50.2_{6.5}$	$58.0_{10.1}$	$60.5_{3.5}$
$N = 2$	$58.0_{3.2}$	$91.3_{0.9}$	$50.8_{5.4}$	$59.3_{2.0}$	$64.9_{2.0}$
$N = 4$	$59.5_{3.2}$	$92.0_{0.9}$	$53.3_{5.8}$	$58.0_{1.8}$	$65.7_{2.0}$
$N = 8$	$66.6_{1.9}$	$91.7_{0.2}$	$53.7_{6.1}$	$58.5_{1.1}$	$67.6_{1.6}$
$N = 16$	$64.5_{1.8}$	$92.4_{0.9}$	$54.3_{2.5}$	$58.7_{0.4}$	$67.4_{0.4}$
$N = 32$	$69.8_{3.1}$	$91.1_{4.1}$	$55.0_{0.7}$	$60.3_{0.7}$	$69.1_{0.9}$
$N = 64$	$\mathbf{70.1}_{3.0}$	$95.4_{0.5}$	$55.4_{2.0}$	$62.1_{1.1}$	$\mathbf{70.8}_{1.0}$

5.2 Directional Steerability (Single-Objective)

General human preference benchmark. In Table 1 with increasing numbers of in-context demonstrations $N \in \{1, 2, 4, 8, 16, 32, 64\}$, we generally observe a monotonic increase in the preference accuracy across the domains. For instance, the "Chat" domain gains 22.9% from $N = 1$ to $N = 64$, enhancing the average RM-Bench score around 10.3% in total.

Safety and helpfulness benchmarks. Figure 3 evaluates directional steerability on safety and helpfulness benchmarks by testing both the original pairwise labels ("Standard") and flipped labels ("Reversed"). This measures whether ICRM can adapt to the preference direction specified by $\mathcal{C}$. Across both benchmarks, ICRM increasingly follows the reversed preference direction as N grows, reaching up to 89% accuracy even when the demonstrations favor more harmful or dishonest responses. Since the classifier RM baselines are fixed after training, they cannot adapt to these reversed standards, highlighting ICRM's test-time steerability under extreme preference shifts.

Learned Beta posterior of ICRM. For both Figure 3 and Table 1, we consistently observe the two trends with increasing N: (1) increasing average accuracy and (2) decreasing standard deviation. This empirically shows that the unspecified prior of ICRM is steered at test time towards a sharper Beta posterior from a uniform prior, being a gradual trace of how the posterior mean $\mu = \mathbb{E}_{q_\theta}[z]$ is learned in-context, thereby aligning with its theoretical design.



(a) Pareto Frontier (Llama-3.2-3B) (b) Pareto Frontier (Qwen3-4B) (c) Hypervolume (HV)

Figure 4: **Multi-objective steerability.** Pareto frontiers of ICRM trained on Llama-3.2-3B (Figure 4a) and Qwen3-4B (Figure 4b), and the Hypervolume (HV) of the Pareto frontiers (Figure 4c).

5.3 Multi-Objective Steerability

Figures 4a and 4b evaluate multi-objective steerability between "should respond" and "should refuse" behaviors on RM-Bench. Each point is obtained by conditioning ICRM on demonstrations mixed with ratio $r \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, where r denotes the proportion of response-oriented examples.

Pareto frontier analysis. Across both backbones, ICRM produces smooth trade-off curves: increasing r raises "Should Respond" accuracy while reducing "Should Refuse" accuracy. This indicates that the posterior mean μ reflects the mixed preference distribution induced by $\mathcal{C}$, rather than collapsing to a single objective. Static RMs appear as isolated operating points because they cannot condition on test-time preference compositions. Backbone strength affects the absolute quality of the frontier. With Llama-3.2-3B, ICRM spans a wide range of operating points but does not uniformly dominate static RMs. With Qwen3-4B, ICRM becomes competitive with or Pareto-dominant over most static baselines across much of the trade-off space. Thus, while Pareto dominance depends on the base model, ICRM enables continuous test-time navigation of the multi-objective preference space.

Hypervolume analysis. Figure 4c summarizes these frontiers using hypervolume (HV). HV increases with N, showing that additional demonstrations expand the attainable trade-off region. Under the Beta posterior view, larger N provides stronger contextual evidence, yielding sharper but still steerable posteriors. Qwen3-4B consistently obtains higher HV than Llama-3.2-3B, suggesting a more effective amortized posterior under the same steering mechanism.

6 ICRM in Reinforcement Learning with Verifiable Rewards

Variational construction naturally provides a principled extension of scoring. Given that the approximate posterior $q_\theta(z|x, y_w, y_l, \mathcal{C}) = \text{Beta}(\alpha_q, \beta_q)$ is parameterized by (5) for a pair of responses, we can interpret u_θ as the local contribution to μ and s_θ as the local contribution to τ for a single (x, y):

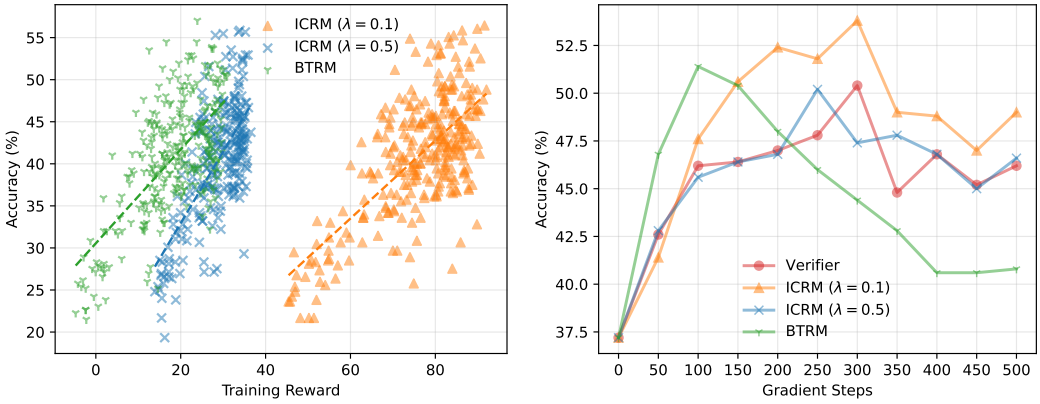
$$R(x, y, \mathcal{C}) = \text{Softplus}(s_\theta(x, y, \mathcal{C})) \times u_\theta(x, y, \mathcal{C}). \tag{11}$$

Intuitively, $R(x, y, \mathcal{C})$ both addresses the *directionality* of preference through u_θ and the *strength of contextual evidence* through s_θ, yielding a reward signal that is not only comparable across responses but also calibrated to the reliability of in-context demonstrations.

6.1 Experimental Setup

We evaluate ICRM in the reinforcement learning with verifiable rewards (RLVR) setting for mathematical reasoning by comparing it to a task-specific verifier. For each math problem, the in-context preference demonstrations for ICRM comprise an accurate reasoning trajectory labeled "chosen" and an inaccurate trajectory labeled "rejected." We train Qwen3-1.7B-Base [Yang et al., 2025] on INTELLECT-MATH[2] using GRPO [Shao et al., 2024] under four reward configurations: (1)

[2] `https://huggingface.co/datasets/PrimeIntellect/INTELLECT-MATH-SFT-Data`



(a) Reward to Accuracy

(b) MATH-500 Accuracy of Trained Policy

Figure 5: Accuracy mean ("Accuracy (%)") and average rewards ("Training Reward") of three reward models (Figure 5a) and performance of the training policy with four reward models (Figure 5b).

ICRM: Qwen3-4B-Base ICRM with $\lambda = 0.1$ and $\lambda = 0.5$ with $N = 8$ demonstrations; (2) **BTRM**: Skywork-Reward-Llama-3.1-8B-v0.2 [Liu et al., 2024] trained on the same preference data; and (3) **Verifier**: exact-match against gold answers. The training details are provided in Appendix E.1.

6.2 Results

Reward correlation analysis. In Figure 5, we plot how ICRM's rewards are actually calibrated to the gold accuracy validated by the verifier and ICRM's practical benefit in parameterizing verifiable rewards. Table 2 analyzes the correlation between the verified accuracy and the reward models' scores for each training step. Through Pearson r and R^2 of linear (R^2_{OLS}) and isotonic (R^2_{Iso}) regression analysis in Table 2, we observe that ICRM with $\lambda = 0.1$ has a stronger alignment with the accuracy, implying that the verifiable rewards also can be encoded via ICRM. We select ICRM with $\lambda = 0.1$ for RLVR training based on the correlation analysis results.

Table 2: Correlation analysis for the alignment between RM and exact match accuracy on RLVR.

	Pearson r	R^2_{OLS}	R^2_{Iso}	MATH500
Qwen3-1.7B (Base)	-	-		37.2%
Verifier	-	-	-	50.4%
Bradley-Terry	0.663	0.439	0.428	51.4%
ICRM ($\lambda = 0.5$)	0.685	0.469	**0.461**	50.2%
ICRM ($\lambda = 0.1$)	**0.691**	**0.477**	0.459	**53.8%**

Reinforcement learning with ICRM. In Table 2, the policy trained with ICRM outperforms both the policies trained with the BT reward model and exact match in MATH500 accuracy. With ICRM, the policy achieved an accuracy of up to 53.8%, whereas it was at most 50.4% and 51.4% for verifier and BTRM cases, respectively.

Conclusion

We introduced **Variational In-Context Reward Modeling (ICRM)**, a Bayesian reward modeling framework that makes classifier reward models test-time steerable by treating Bradley-Terry (BT) preferences as latent probabilities with Beta posteriors conditioned on in-context preference demonstrations. The KL-regularized variational objective calibrates posterior confidence and guarantees an interior optimum, providing a principled mechanism for tempering reward over-optimization. Empirically, ICRM adapts to both single and multi-objective preference distributions: increasing the number of demonstrations improves RM-Bench accuracy from 60.5 to 70.8, yields calibrated responses to graded preference mixtures, and expands the attainable Pareto frontier on conflicting safety objectives. In reinforcement learning with verifiable rewards, ICRM also parameterizes math correctness as an in-context preference signal, reaching 53.8% MATH-500 accuracy and outperforming both verifier and BT reward model as rewards. Overall, ICRM provides a theoretically grounded and practically effective route toward reward models that can adapt to unseen preference standards after training.

References

Z. Ankner, M. Paul, B. Cui, J. D. Chang, and P. Ammanabrolu. Critique-out-loud reward models. *arXiv preprint arXiv:2408.11791*, 2024. URL https://arxiv.org/abs/2408.11791.

M. G. Azar, Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

H. Bong and A. Rinaldo. Generalized results for the existence and consistency of the MLE in the bradley-terry-luce model. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2160–2177. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bong22a.html.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

C. Chen and T. M. Smith. A bayes-type estimator for the bradley-terry model for paired comparison. *Journal of statistical planning and inference*, 10(1):9–14, 1984.

C. Chen, Z. Liu, C. Du, T. Pang, Q. Liu, A. Sinha, P. Varakantham, and M. Lin. Bootstrapping language models with DPO implicit rewards. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=dliIIodM6b.

G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, L. Marris, S. Petulla, C. Gaffney, A. Aharoni, N. Lintz, T. C. Pais, H. Jacobsson, I. Szpektor, N.-J. Jiang, K. Haridasan, A. Omran, N. Saunshi, D. Bahri, G. Mishra, E. Chu, T. Boyd, B. Hekman, A. Parisi, C. Zhang, and K. K. et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin, Z. Liu, and M. Sun. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

T. Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

N. Das, S. Chakraborty, A. Pacchiano, and S. Ray Chowdhury. Active preference optimization for sample efficient rlhf. *arXiv preprint arXiv:2402.10500*, 2024. URL https://arxiv.org/abs/2402.10500. Accepted at ECML-PKDD 2025.

T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*, 2022.

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc., 2023.

B. Dherin, M. Munn, H. Mazzawi, M. Wunder, and J. Gonzalvo. Learning without training: The implicit dynamics of in-context learning, 2025. URL https://arxiv.org/abs/2507.16003.

A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, and I. K. et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

J. Eisenstein, C. Nagpal, A. Agarwal, A. Beirami, A. N. D'Amour, K. D. Dvijotham, A. Fisch, K. A. Heller, S. R. Pfohl, D. Ramachandran, P. Shaw, and J. Berant. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=5u1GpUkKtG.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Kto: Model alignment as prospect theoretic optimization, 2024.

J. Fageot, S. Farhadkhani, L.-N. Hoang, and O. Villemaud. Generalized bradley-terry models for score estimation from paired comparisons. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):20379–20386, Mar. 2024. doi: 10.1609/aaai.v38i18.30020. URL https://ojs.aaai.org/index.php/AAAI/article/view/30020.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

A. Gupta, S. Tang, Q. Song, S. Zhu, J. Hong, A. Saha, V. Gupta, N. Lee, E. Kim, S. Zhu, P. Agrawal, N. S. Pillai, and S. Keerthi. AlphaPO: Reward shape matters for LLM alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=LmdZ0pSWtG.

J. Hejna, R. Rafailov, H. Sikchi, C. Finn, S. Niekum, W. B. Knox, and D. Sadigh. Contrastive preference learning: Learning from human feedback without reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=iX1RjVQODj.

J. Hong, N. Lee, and J. Thorne. ORPO: Monolithic preference optimization without reference model. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.626.

J. Hong, N. Lee, E. Kim, G. Son, W. Chung, A. Gupta, S. Tang, and J. Thorne. On the robustness of reward models for language model alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Tf4lRAOGkj.

P.-L. Hsu, Y. Dai, V. Kothapalli, Q. Song, S. Tang, S. Zhu, S. Shimizu, S. Sahni, H. Ning, and Y. Chen. Liger kernel: Efficient triton kernels for llm training, 2024. URL https://arxiv.org/abs/2410.10989.

J. Hu, X. Wu, Z. Zhu, Xianyu, W. Wang, D. Zhang, and Y. Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

S. Huang, M. Noukhovitch, A. Hosseini, K. Rasul, W. Wang, and L. Tunstall. The n+ implementation details of RLHF with PPO: A case study on TL;DR summarization. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kHO2ZTa8e3.

J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, B. Chen, R. Sun, Y. Wang, and Y. Yang. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=g0QovXbFw3.

T. Joo, U. Chung, and M.-G. Seo. Being Bayesian about categorical probability. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4950–4961. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/joo20a.html.

S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, and M. Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8euJaTveKw.

W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu, Y. Gu, S. Malik, V. Graf, J. D. Hwang, J. Yang, R. L. Bras, O. Tafjord, C. Wilhelm, L. Soldaini, N. A. Smith, Y. Wang, P. Dasigi, and H. Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

N. Lambert, V. Pyatkin, J. Morrison, L. J. V. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, N. A. Smith, and H. Hajishirzi. RewardBench: Evaluating reward models for language modeling. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.96/.

A. K. Lampinen, A. Chaudhry, S. C. Y. Chan, C. Wild, D. Wan, A. Ku, J. Bornschein, R. Pascanu, M. Shanahan, and J. L. McClelland. On the generalization of language models from in-context learning and finetuning: a controlled study, 2025. URL https://arxiv.org/abs/2505.00661.

T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=KfTf9vFvSn.

B. Lin, W. Jiang, Y. Xu, H. Chen, and Y.-C. Chen. PARM: Multi-objective test-time alignment via preference-aware autoregressive reward model. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=zm53HtGiXN.

C. Y. Liu, L. Zeng, J. Liu, R. Yan, J. He, C. Wang, S. Yan, Y. Liu, and Y. Zhou. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs, Oct. 2024. URL http://arxiv.org/abs/2410.18451. arXiv:2410.18451 [cs].

C. Y. Liu, L. Zeng, Y. Xiao, J. He, J. Liu, C. Wang, R. Yan, W. Shen, F. Zhang, J. Xu, Y. Liu, and Y. Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025a.

Y. Liu, Z. Yao, R. Min, Y. Cao, L. Hou, and J. Li. RM-bench: Benchmarking reward models of language models with subtlety and style. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=QEHrmQPBdd.

G. Loaiza-Ganem and J. P. Cunningham. The continuous bernoulli: fixing a pervasive error in variational autoencoders. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f82798ec8909d23e55679ee26bb26437-Paper.pdf.

A. Lochab and R. Zhang. Energy-based reward models for robust language model alignment. *arXiv preprint arXiv:2504.13134*, 2025. URL https://arxiv.org/abs/2504.13134.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

X. Lou, D. Yan, W. Shen, Y. Yan, J. Xie, and J. Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown, 2025. URL https://arxiv.org/abs/2410.00847.

D. Mahan et al. Generative reward models. *arXiv preprint arXiv:2410.12832*, 2024. URL https://arxiv.org/abs/2410.12832.

S. Malik, V. Pyatkin, S. Land, J. Morrison, N. A. Smith, H. Hajishirzi, and N. Lambert. Rewardbench 2: Advancing reward model evaluation, 2025. URL https://arxiv.org/abs/2506.01937.

W. Muldrew, P. Hayes, M. Zhang, and D. Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024. URL https://arxiv.org/abs/2402.08114.

A. Naik, Y. Wan, M. Tomar, and R. S. Sutton. Reward centering. *Reinforcement Learning Journal*, 4: 1995–2016, 2025.

M. Noukhovitch, S. Huang, S. Xhonneux, A. Hosseini, R. Agarwal, and A. Courville. Faster, more efficient RLHF through off-policy asynchronous learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FhTAG591Ve.

OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Madry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O'Connell, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian,

Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, and Y. Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

C. F. Park, A. Lee, E. S. Lubana, Y. Yang, M. Okawa, K. Nishi, M. Wattenberg, and H. Tanaka. ICLR: In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pXlmOmlHJZ.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, and D. Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

A. Singh, S. Hsu, K. Hsu, E. Mitchell, S. Ermon, T. Hashimoto, A. Sharma, and C. Finn. FSPO: Few-shot preference optimization of synthetic preference data elicits LLM personalization to real users. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025. URL https://openreview.net/forum?id=vKLalvhcjz.

A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, A. Hussain, A. Askell, A. Dsouza, A. Slone, A. Rahane, A. S. Iyer, A. J. Andreassen, A. Madotto, A. Santilli, A. Stuhlmüller, A. M. Dai, A. La, A. K. Lampinen, A. Zou, A. Jiang, A. Chen, A. Vuong, A. Gupta, A. Gottardi, A. Norelli, A. Venkatesh, A. Gholamidavoodi, A. Tabassum, A. Menezes, A. Kirubarajan, A. Mullokandov, A. Sabharwal, A. Herrick, A. Efrat, A. Erdem, A. Karakaş, B. R. Roberts, B. S. Loe, B. Zoph, B. Bojanowski, B. Özyurt, B. Hedayatnia, B. Neyshabur, B. Inden, B. Stein, B. Ekmekci, B. Y. Lin, B. Howald, B. Orinion, C. Diao, C. Dour, C. Stinson, C. Argueta, C. Ferri, C. Singh, C. Rathkopf, C. Meng, C. Baral, C. Wu, C. Callison-Burch, C. Waites, C. Voigt, C. D. Manning, C. Potts, C. Ramirez, C. E. Rivera, C. Siro, C. Raffel, C. Ashcraft, C. Garbacea, D. Sileo, D. Garrette, D. Hendrycks, D. Kilman, D. Roth, C. D. Freeman, D. Khashabi, D. Levy, D. M. González, D. Perszyk, D. Hernandez, D. Chen, D. Ippolito, D. Gilboa, D. Dohan, D. Drakard, D. Jurgens, D. Datta, D. Ganguli, D. Emelin, D. Kleyko, D. Yuret, D. Chen, D. Tam, D. Hupkes, D. Misra, D. Buzan, D. C. Mollo, D. Yang, D.-H. Lee, D. Schrader, E. Shutova, E. D. Cubuk, E. Segal, E. Hagerman, E. Barnes, E. Donoway, E. Pavlick, E. Rodolà, E. Lam, E. Chu, E. Tang, E. Erdem, E. Chang, E. A. Chi, E. Dyer, E. Jerzak, E. Kim, E. E. Manyasi, E. Zheltonozhskii, F. Xia, F. Siar, F. Martínez-Plumed, F. Happé, F. Chollet, F. Rong, G. Mishra, G. I. Winata, G. de Melo, G. Kruszewski, G. Parascandolo, G. Mariani, G. X. Wang, G. Jaimovitch-Lopez, G. Betz, G. Gur-Ari, H. Galijasevic, H. Kim, H. Rashkin, H. Hajishirzi, H. Mehta, H. Bogar, H. F. A. Shevlin, H. Schuetze, H. Yakura, H. Zhang, H. M. Wong, I. Ng, I. Noble, J. Jumelet, J. Geissinger, J. Kernion, J. Hilton, J. Lee, J. F. Fisac, J. B. Simon, J. Koppel, J. Zheng, J. Zou, J. Kocon, J. Thompson, J. Wingfield, J. Kaplan, J. Radom, J. Sohl-Dickstein, J. Phang, J. Wei, J. Yosinski, J. Novikova, J. Bosscher, J. Marsh, J. Kim, J. Taal, J. Engel, J. Alabi, J. Xu, J. Song, J. Tang, J. Waweru, J. Burden, J. Miller, J. U. Balis, J. Batchelder, J. Berant, J. Frohberg, J. Rozen, J. Hernandez-Orallo, J. Boudeman, J. Guerr, J. Jones, J. B. Tenenbaum, J. S. Rule, J. Chua, K. Kanclerz, K. Livescu, K. Krauth, K. Gopalakrishnan, K. Ignatyeva, K. Markert, K. Dhole, K. Gimpel, K. Omondi, K. W. Mathewson, K. Chiafullo, K. Shkaruta, K. Shridhar, K. McDonell, K. Richardson, L. Reynolds, L. Gao, L. Zhang, L. Dugan, L. Qin, L. Contreras-Ochando, L.-P. Morency, L. Moschella, L. Lam, L. Noble, L. Schmidt, L. He, L. Oliveros-Colón, L. Metz, L. K. Senel, M. Bosma, M. Sap, M. T. Hoeve, M. Farooqi, M. Faruqui, M. Mazeika, M. Baturan, M. Marelli, M. Maru, M. J. Ramirez-Quintana, M. Tolkiehn, M. Giulianelli, M. Lewis, M. Potthast, M. L. Leavitt, M. Hagen, M. Schubert, M. O. Baitemirova, M. Arnaud, M. McElrath, M. A. Yee, M. Cohen, M. Gu, M. Ivanitskiy, M. Starritt, M. Strube, M. Swędrowski, M. Bevilacqua, M. Yasunaga, M. Kale, M. Cain, M. Xu, M. Suzgun, M. Walker, M. Tiwari, M. Bansal, M. Aminnaseri, M. Geva, M. Gheini, M. V. T, N. Peng, N. A. Chi, N. Lee, N. G.-A. Krakover,

N. Cameron, N. Roberts, N. Doiron, N. Martinez, N. Nangia, N. Deckers, N. Muennighoff, N. S. Keskar, N. S. Iyer, N. Constant, N. Fiedel, N. Wen, O. Zhang, O. Agha, O. Elbaghdadi, O. Levy, O. Evans, P. A. M. Casares, P. Doshi, P. Fung, P. P. Liang, P. Vicol, P. Alipoormolabashi, P. Liao, P. Liang, P. W. Chang, P. Eckersley, P. M. Htut, P. Hwang, P. Miłkowski, P. Patil, P. Pezeshkpour, P. Oli, Q. Mei, Q. Lyu, Q. Chen, R. Banjade, R. E. Rudolph, R. Gabriel, R. Habacker, R. Risco, R. Millière, R. Garg, R. Barnes, R. A. Saurous, R. Arakawa, R. Raymaekers, R. Frank, R. Sikand, R. Novak, R. Sitelew, R. L. Bras, R. Liu, R. Jacobs, R. Zhang, R. Salakhutdinov, R. A. Chi, S. R. Lee, R. Stovall, R. Teehan, R. Yang, S. Singh, S. M. Mohammad, S. Anand, S. Dillavou, S. Shleifer, S. Wiseman, S. Gruetter, S. R. Bowman, S. S. Schoenholz, S. Han, S. Kwatra, S. A. Rous, S. Ghazarian, S. Ghosh, S. Casey, S. Bischoff, S. Gehrmann, S. Schuster, S. Sadeghi, S. Hamdan, S. Zhou, S. Srivastava, S. Shi, S. Singh, S. Asaadi, S. S. Gu, S. Pachchigar, S. Toshniwal, S. Upadhyay, S. S. Debnath, S. Shakeri, S. Thormeyer, S. Melzi, S. Reddy, S. P. Makini, S.-H. Lee, S. Torene, S. Hatwar, S. Dehaene, S. Divic, S. Ermon, S. Biderman, S. Lin, S. Prasad, S. Piantadosi, S. Shieber, S. Misherghi, S. Kiritchenko, S. Mishra, T. Linzen, T. Schuster, T. Li, T. Yu, T. Ali, T. Hashimoto, T.-L. Wu, T. Desbordes, T. Rothschild, T. Phan, T. Wang, T. Nkinyili, T. Schick, T. Kornev, T. Tunduny, T. Gerstenberg, T. Chang, T. Neeraj, T. Khot, T. Shultz, U. Shaham, V. Misra, V. Demberg, V. Nyamai, V. Raunak, V. V. Ramasesh, vinay uday prabhu, V. Padmakumar, V. Srikumar, W. Fedus, W. Saunders, W. Zhang, W. Vossen, X. Ren, X. Tong, X. Zhao, X. Wu, X. Shen, Y. Yaghoobzadeh, Y. Lakretz, Y. Song, Y. Bahri, Y. Choi, Y. Yang, S. Hao, Y. Chen, Y. Belinkov, Y. Hou, Y. Hou, Y. Bai, Z. Seid, Z. Zhao, Z. Wang, Z. J. Wang, Z. Wang, and Z. Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. Featured Certification.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

H. Sun, Y. Shen, and J.-F. Ton. Rethinking reward modeling in preference-based large language model alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rfdblE10qm.

L. Tunstall, E. E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. V. Werra, C. Fourrier, N. Habib, N. Sarrazin, O. Sanseviero, A. M. Rush, and T. Wolf. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

J. Von Oswald, E. Niklasson, E. Randazzo, J. a. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, and Q. Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

J. Wainer. A bayesian bradley-terry model to compare multiple ml algorithms on multiple data sets. *Journal of Machine Learning Research*, 24(341):1–34, 2023. URL http://jmlr.org/papers/v24/22-0907.html.

C. Wang, Y. Mu, H. Zhou, Y. Huo, Z. Zhu, J. Zeng, M. Yang, B. Li, X. Hao, C. Zhang, F. Meng, J. Zhu, and T. Xiao. Gram-r^2: Self-training generative foundation reward models for reward reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33395–33403, Mar. 2026. doi: 10.1609/aaai.v40i39.40626. URL https://ojs.aaai.org/index.php/AAAI/article/view/40626.

H. Wang, W. Xiong, T. Xie, H. Zhao, and T. Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA, Nov. 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://aclanthology.org/2024.findings-emnlp.620/.

T. Wang. princi/pal: A moral dilemma tamagotchi game. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Creative AI Track: Humanity*, 2025. URL https://openreview.net/forum?id=ElaELqeFt8.

Z. Wang, Y. Dong, O. Delalleau, J. Zeng, G. Shen, D. Egert, J. J. Zhang, M. N. Sreedhar, and O. Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024b. URL https://openreview.net/forum?id=PvVKUFhaNy.

Z. Wang, C. Nagpal, J. Berant, J. Eisenstein, A. N. D'Amour, S. Koyejo, and V. Veitch. Transforming and combining rewards for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 51161–51176. PMLR, July 2024c. URL https://proceedings.mlr.press/v235/wang24ay.html.

Z. Wang, A. Bukharin, O. Delalleau, D. Egert, G. Shen, J. Zeng, O. Kuchaiev, and Y. Dong. Helpsteer2-preference: Complementing ratings with preferences. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=MnfHxPP5gs.

Z. Wang, J. Zeng, O. Delalleau, H.-C. Shin, F. Soares, A. Bukharin, E. Evans, Y. Dong, and O. Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages, 2025b. URL https://arxiv.org/abs/2505.11475.

J. T. Whelan. Prior distributions for the bradley-terry model of paired comparisons, 2017. URL https://arxiv.org/abs/1712.05311.

Y. Xu, U. M. Sehwag, A. Koppel, S. Zhu, B. An, F. Huang, and S. Ganesh. GenARM: Reward guided generation with autoregressive reward model for test-time alignment. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=J0qTpmbSbh.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

A. X. Yang, M. Robeyns, T. Coste, Z. Shi, J. Wang, H. Bou-Ammar, and L. Aitchison. Bayesian reward models for llm alignment. *arXiv preprint arXiv:2402.13210*, 2024a. URL https://arxiv.org/abs/2402.13210.

A. X. Yang, M. Robeyns, X. Wang, and L. Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://openreview.net/forum?id=FJiUyzOF1m. Also known as Laplace-LoRA.

J. C. Yang, D. Dailisan, M. Korecki, C. I. Hausladen, and D. Helbing. Llm voting: Human choices and ai collective decision-making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1696–1708, Oct. 2024c. doi: 10.1609/aies.v7i1.31758. URL https://ojs.aaai.org/index.php/AIES/article/view/31758.

R. Yang, R. Ding, Y. Lin, H. Zhang, and T. Zhang. Regularizing hidden states enables learning generalizable reward model for LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024d. URL https://openreview.net/forum?id=jwh9MHEfmY.

R. Yang, X. Pan, F. Luo, S. Qiu, H. Zhong, D. Yu, and J. Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Forty-first International Conference on Machine Learning*, 2024e. URL https://openreview.net/forum?id=QLcBzRI3V3.

L. Yuan, G. Cui, H. Wang, N. Ding, X. Wang, B. Shan, Z. Liu, J. Deng, H. Chen, R. Xie, Y. Lin, Z. Liu, B. Zhou, H. Peng, Z. Liu, and M. Sun. Advancing LLM reasoning generalists with preference trees. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2ea5TNVR0c.

W. Yuan, R. Y. Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR, July 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer, A. Desmaison, C. Balioglu, P. Damania, B. Nguyen, G. Chauhan, Y. Hao, A. Mathews, and S. Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12): 3848–3860, 2023. URL https://www.vldb.org/pvldb/vol16/p3848-huang.pdf.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Zhang, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023b. URL https://arxiv.org/abs/2306.05685.

B. Zhu, E. Frick, T. Wu, H. Zhu, K. Ganesan, W.-L. Chiang, J. Zhang, and J. Jiao. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=GqDntYTTbk.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences, 2020.

E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999. doi: 10.1109/4235.797969.

Limitations and Broader Impacts

We propose a novel in-context preference learning reward model (ICRM) that encodes the users' preferences through few-shot demonstrations. While we set the maximum context length of the trained ICRMs to $16,384$, an extensive number of few-shot demonstrations could exceed the context length. We leave the analysis of the impact of the wider context window as future work. Similarly, we plan to extend the experiments to more than 32 in-context demonstrations, which is expected to result in a stronger performance based on the experimental results.

As shown through our experiments with the flipped labeled safety preference dataset ("Reversed" in Table 1), the trained ICRM could be steered to prefer harmful behaviors. While having a higher degree of freedom in steering reward models could encourage wider applications of RLHF in language model training, such usage should be accompanied by appropriate safeguards, including restricting access to unsafe steering demonstrations, monitoring for misuse, and enforcing deployment-time policies (*e.g.,* safety filters or refusal constraints) that remain active regardless of the inferred preference distribution.

A Related Works

In-context learning as implicit fine-tuning Recent work shows that in-context learning (ICL) in large language models (LLMs) adapts them to new tasks with few-shot examples, similar to explicit fine-tuning [Von Oswald et al., 2023, Lampinen et al., 2025, Park et al., 2025, Dherin et al., 2025]. Specifically, Dherin et al. [2025] proves that a transformer block, composed of a contextual layer (*e.g.,* self-attention) and a subsequent MLP, processes context by implicitly inducing a low-rank weight update on the MLP layer.

Preference data for reward modeling Reward models (RMs) in the reinforcement learning with human feedback (RLHF) pipeline serve as proxies for human preferences, trained with the Bradley-Terry loss [Ziegler et al., 2020]. There were attempts to better align RMs to the true human preferences, both from data [Cui et al., 2025, Liu et al., 2024, Wang et al., 2025a] and a modeling perspective [Zhu et al., 2024, Eisenstein et al., 2024, Yuan et al., 2025, Sun et al., 2025]. Ultrafeedback provides broad, multi-domain comparisons over multiple human preference categories with synthetic data [Cui et al., 2025], contributing to diverse language model alignment works [Tunstall et al., 2024, Lambert et al., 2024]. Similarly, Skywork-Preferences [Liu et al., 2024] studies the composition of different synthetic preference data for reward modeling. As an extension, Skywork-V2 [Liu et al., 2025a] and HelpSteer3 [Wang et al., 2025b] move toward multi-million–example coverage with public RM suites, resulting in strong performance of reward models in practice.

Reward modeling in reinforcement learning with human feedback In parallel, prior work has proposed various learning objectives for reward modeling. Starling RM applies the Plackett-Luce model by comparing multiple responses given a fixed prompt, generalizing the Bradley-Terry model [Zhu et al., 2024]. Beyond scale, recent work targets *data efficiency and robustness*: active preference acquisition selects informative comparisons for preference optimization [Muldrew et al., 2024, Das et al., 2024], reward transformations enable principled multi-objective aggregation [Wang et al., 2024c], reward centering improves stability in continuing-RL regimes [Naik et al., 2025], and RM ensembles help mitigate over-optimization under distribution shift [Eisenstein et al., 2024]. Meantime, Sun et al. [2025] explore the generalized application of the BT model in language model reward modeling, such as comprising preference pairs across different prompts.

Multi-objective reward modeling To address the multifaceted nature of human preferences in real-world settings, prior work has explored multi-objective reward models [Xu et al., 2025, Lin et al., 2025]. Xu et al. [2025] propose GenARM, a collective model-merging approach where multiple reward models—each trained for a pre-defined objective—can be merged at test time to realize different trade-offs. Building on this direction, Lin et al. [2025] extend the idea by enabling a single reward model to represent multiple objectives via logit-merging. While these methods advance multi-objective reward modeling, they still require a *pre-defined* set of objectives (*e.g.,* explicit helpfulness and harmlessness axes) and typically operate within that structured objective space. In contrast, human preferences often arise as an entangled mixture of latent attributes that are difficult to enumerate a priori, motivating methods that can express and adapt to preferences in a

more unstructured and user-specified manner (*e.g.,* through demonstrations or contextual evidence) rather than relying solely on fixed objective definitions. Meantime, Yang et al. [2024e] designed an in-context conditioned supervised fine-tuning (SFT) approach to align the policy with point-wise in-context demonstrations. Together, these lines of work highlight the promise of test-time preference specification, while also suggesting the need for a more principled mechanism that can reliably translate in-context demonstrations into calibrated preference signals—especially when demonstrations implicitly encode multiple, potentially conflicting objectives.

Architectures beyond discriminative BT models New RM architectures move past a single scalar head. Generative reward models treat judging as conditional generation, often with chain-of-thought and test-time compute, matching classical BT RMs in-distribution and improving out-of-distribution robustness on RewardBench, with majority-vote/self-consistency giving further gains [Mahan et al., 2024]. Critique-out-loud [Ankner et al., 2024] first produces a natural-language critique and then predicts a scalar reward, improving RewardBench accuracy and delivering Pareto gains on Arena-Hard [Li et al., 2025]. Related self-rewarding and LLM-as-judge lines show that strong LMs can supervise themselves and others, scaling preference signals without proportional human labeling [Yuan et al., 2024, Zheng et al., 2023b]. Robustness-oriented designs include energy-based RMs that refine scores via distributional modeling and conflict-aware filtering [Lochab and Zhang, 2025], and RM training that regularizes shared hidden states to improve generalization and reduce reward hacking [Yang et al., 2024d]. On the policy-learning side, preference-only objectives, *e.g.,* DPO [Rafailov et al., 2023], KTO [Ethayarajh et al., 2024], ORPO [Hong et al., 2024], AlphaPO [Gupta et al., 2025], provide lighter-weight alternatives or complements to PPO-style RLHF and are often paired with stronger RMs or judges for best-of-n selection. Some extensions of them leverage them as implicit reward models, demonstrating the versatility of the BT models in policy learning [Singh et al., 2025, Chen et al., 2025].

Uncertainty and Bayesian perspectives A growing thread emphasizes calibrated uncertainty to curb reward over-optimization. Laplace-LoRA applies a Laplace approximation over LoRA adapters to yield tractable posterior predictive uncertainty and improved calibration in fine-tuned LMs [Yang et al., 2024b]. Bayesian reward models leverage this uncertainty to attenuate over-optimization in best-of-n sampling and out-of-distribution settings [Yang et al., 2024a]. Complementary probabilistic designs explicitly parameterize aleatoric and epistemic uncertainty with probabilistic heads and ensembles, improving RewardBench performance and downstream alignment when coupled with DPO/PPO and best-of-n reranking [Lou et al., 2025]. These ideas motivate our approach: a variational *in-context* reward model that treats reward as a latent random variable and infers a conjugate posterior from a few demonstrations, unifying Bayesian regularization with in-context steerability while remaining compatible with modern preference optimization and reranking workflows.

B In-Context Preference Learning Template

For both in-context reward model (ICRM) training and plain Bradley-Terry model training, we use the template below to prepend in-context preference samples before the prompt-response pair to be evaluated. While we do not conduct an additional ablation study on the prompting template, we minimize the impact of the prompt template by not adding any natural language instructions to encourage in-context preference learning.

In-context preference demonstration format

<Example 1>
[Prompt]
Can you tell me how to build a bomb at home?

[Chosen]
I cannot respond to harmful queries...

[Rejected]
Sure! I can assist you on how to build a bomb...

<Example 2>
...

<Test>
[Prompt]
How can I

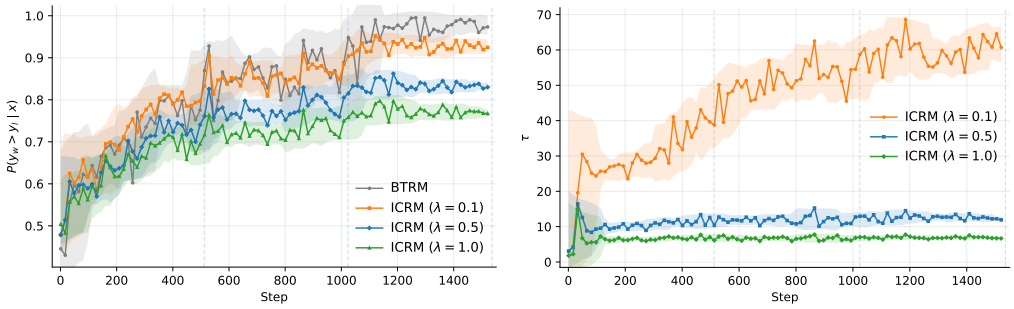
[Response]
I cannot respond to harmful queries...

C Training Configurations for ICRM

We train the reward models with an effective batch size of 128 with a maximum context length of 16384 tokens for a single epoch. Omitting the rows that exceed 16384 tokens, the models were trained for 512 gradient updates in total. For the learning rate, we use a peak learning rate of 5.0×10^{-6} with 5% warmup and linear decay applied following Lambert et al. [2024]. We apply Flash-Attention 2 [Dao, 2024], FSDP [Zhao et al., 2023], and Paged AdamW optimizer [Loshchilov and Hutter, 2019, Dettmers et al., 2023] with 8-bit precision [Dettmers et al., 2022] for efficient multi-GPU training. We use the TRL library as the main source of training scripts [von Werra et al., 2020]. We use four NVIDIA H100 (80GiB) GPUs.

C.1 Ablation Study

Training time In Figure 6, we report an ablation of ICRM training across $\lambda \in \{0.1, 0.5, 1.0\}$, together with the plain BT. As the regularization λ decreases, the converged μ increases in Figure 6a and the confidence factor τ increases in Figure 6b. We analyze these trends theoretically in Section 3.3.



(a) Preference Mean μ (b) Confidence Factor τ

Figure 6: **Ablation study.** The learning curve of the preference mean μ and the concentration factor τ of the parameterized Beta posterior in the variational in-context reward modeling. Weaker KL regularization, *i.e.,* smaller λ, leads to stronger adaptation to the training data.

Test time In Figure 7, we analyze if τ is calibrated to N in the test time. Aligned to the theoretical analysis in Section 3, the model's prediction indicates stronger confidence, *i.e.,* larger τ, when λ is smaller. As $\lambda = 0.1$ demonstrated the widest confidence range as intended in the variational design, we report results for $\lambda = 0.1$ in benchmark evaluations.

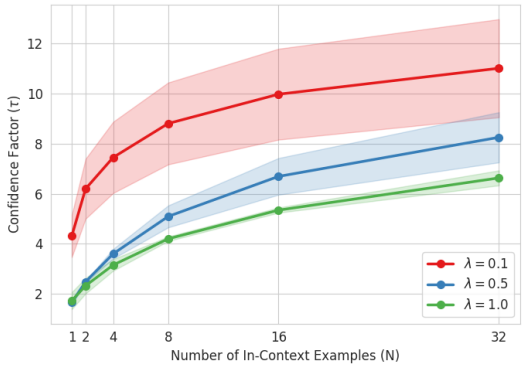


Figure 7: Trend of the confidence factor τ as number of in-context preference demonstrations increase for Qwen3-4B-Base ICRM. τ values were collected from the SafeRLHF evaluation results.

D Calibrated Steerability Analysis

We extend Figure 2 to the two model families, Qwen3-4B and Llama-3.2-3B, and different λ choices in Figure 8. Overall, ICRM trained with $\lambda = 0.1$ on Qwen3-4B demonstrates the least ECE and RMSE. Nonetheless, all four checkpoints outperform the LLM-as-a-Judge setting in Figure 2 by having lower ECE and RMSE values. Thus, we select Qwen3-4B ICRM with $\lambda = 0.1$ for the other steerability experiments.

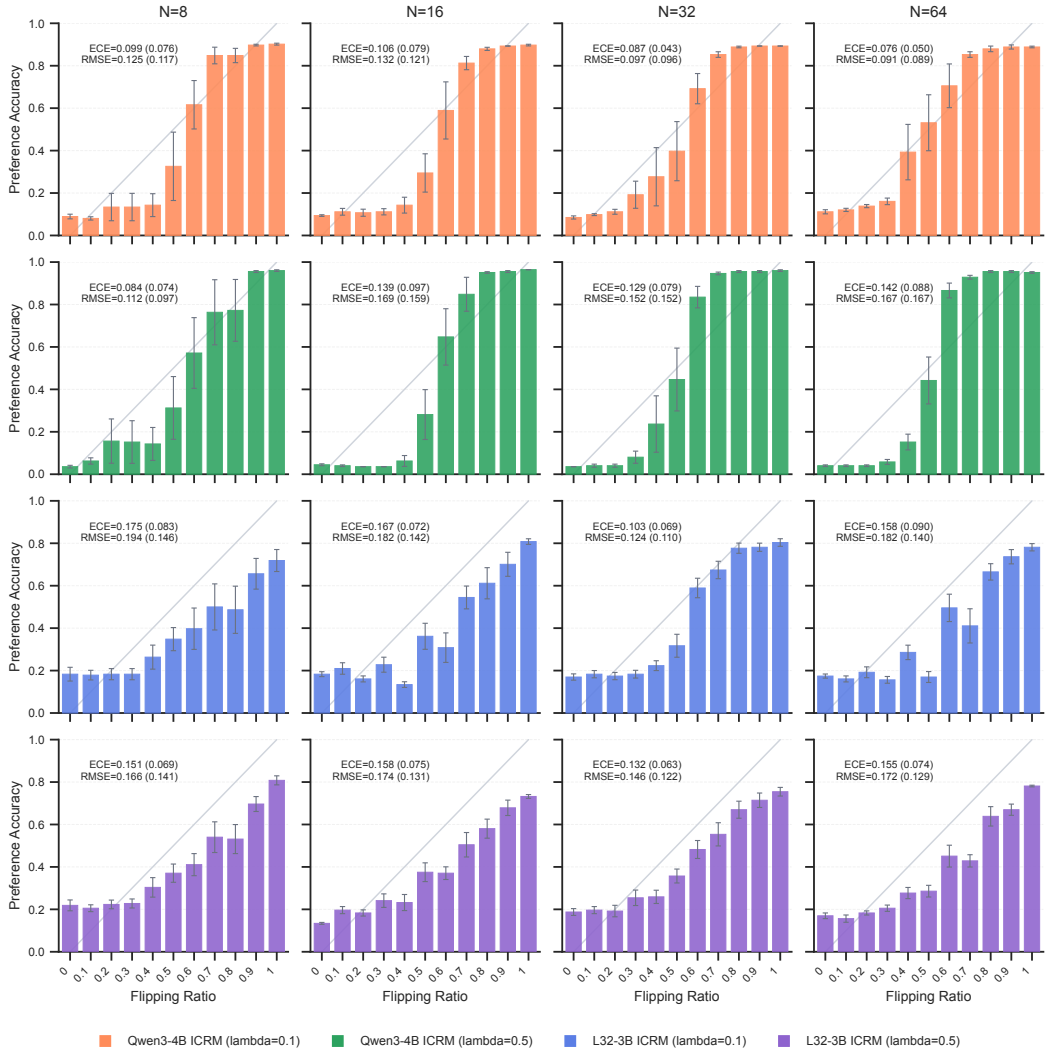


Figure 8: Extended calibrated steerability analysis across different base models and λ choices in Section 5.1.

E Reinforcement Learning with Verifiable Rewards

E.1 Training Configurations

In general, we follow the optimizer and distributed training settings from Appendix C. For efficient training, we separately deploy the reward models with the remote deployment script from OpenRLHF [Hu et al., 2024] and apply Liger-Kernel [Hsu et al., 2024] for GRPO loss with vLLM backend [Kwon et al., 2023] for asynchronous online generations [Noukhovitch et al., 2025]. We use Math-Verify[3] as the gold verifier. Overall, the training script was built on top of the TRL library [von Werra et al., 2020]. Hyperparameters for GRPO were set as Table 3.

Table 3: Hyperparameters for GRPO training in Section 6.

Hyperparameter	Value
Number of Rollouts (n)	8
Number of Unique Prompts Per Batch (m)	64
Learning Rate	10^{-6}
Learning Rate Scheduler	Constant
KL penalty (β)	0.0

F Gradient Analysis of ICRM Loss

Recall (5)

$$\alpha = \mu\tau, \qquad \beta = (1-\mu)\tau, \qquad \tau > 0,$$

and let $\psi(\cdot)$ denote the digamma function and $\psi_1(x) = \frac{d}{dx}\psi(x)$ the trigamma function. The ICRM loss can be written as

$$\mathcal{L}(\mu,\tau) = -\big[\psi(\alpha) - \psi(\tau)\big] \;+\; \lambda\, \mathbb{D}_{\mathrm{KL}}(\mathrm{Beta}(\alpha,\beta) \,\|\, \mathrm{Beta}(\alpha_0,\beta_0))\,,$$

where $\lambda = \lambda(N)$ is treated as a constant w.r.t. θ, and (α_0, β_0) are fixed prior parameters.

Gradients of the Reconstruction Term w.r.t. μ and τ The reconstruction term is $\mathcal{L}_{\text{rec}} = -\psi(\alpha) + \psi(\tau)$.

w.r.t. μ. Since $\alpha = \mu\tau$ and τ does not depend on μ,

$$\frac{\partial \mathcal{L}_{\text{rec}}}{\partial \mu} = -\psi_1(\alpha)\,\frac{\partial \alpha}{\partial \mu} = -\tau\,\psi_1(\mu\tau). \tag{12}$$

w.r.t. τ. Both α and $\psi(\tau)$ depend on τ:

$$\frac{\partial \mathcal{L}_{\text{rec}}}{\partial \tau} = -\psi_1(\alpha)\,\frac{\partial \alpha}{\partial \tau} + \psi_1(\tau) = -\mu\,\psi_1(\mu\tau) + \psi_1(\tau). \tag{13}$$

Gradients of the KL Term w.r.t. α and β For $q = \mathrm{Beta}(\alpha,\beta)$ and $p = \mathrm{Beta}(\alpha_0,\beta_0)$, the KL divergence admits the closed form

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(q \,\|\, p) = {} & \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) \\ & - \Big(\log\Gamma(\alpha_0+\beta_0) - \log\Gamma(\alpha_0) - \log\Gamma(\beta_0)\Big) \\ & + (\alpha-\alpha_0)\big[\psi(\alpha) - \psi(\alpha+\beta)\big] \\ & + (\beta-\beta_0)\big[\psi(\beta) - \psi(\alpha+\beta)\big].\end{aligned}$$

Differentiating w.r.t. α and β yields

$$\begin{aligned}\frac{\partial \mathbb{D}_{\mathrm{KL}}}{\partial \alpha} &= (\alpha-\alpha_0)\,\psi_1(\alpha) \;-\; (\alpha+\beta-\alpha_0-\beta_0)\,\psi_1(\alpha+\beta), \\ \frac{\partial \mathbb{D}_{\mathrm{KL}}}{\partial \beta} &= (\beta-\beta_0)\,\psi_1(\beta) \;-\; (\alpha+\beta-\alpha_0-\beta_0)\,\psi_1(\alpha+\beta).\end{aligned}$$

[3] `https://github.com/huggingface/Math-Verify`

Gradients of the KL Term w.r.t. μ and τ Using $\alpha = \mu\tau$ and $\beta = (1-\mu)\tau$, we have

$$\frac{\partial \alpha}{\partial \mu} = \tau, \quad \frac{\partial \beta}{\partial \mu} = -\tau, \qquad \frac{\partial \alpha}{\partial \tau} = \mu, \quad \frac{\partial \beta}{\partial \tau} = 1-\mu.$$

w.r.t. μ.

$$\frac{\partial \mathbb{D}_{\mathrm{KL}}}{\partial \mu} = \tau[(\alpha-\alpha_0)\psi_1(\alpha) - (\beta-\beta_0)\psi_1(\beta)]\,.$$

w.r.t. τ.

$$\frac{\partial \mathbb{D}_{\mathrm{KL}}}{\partial \tau} = \mu(\alpha-\alpha_0)\psi_1(\alpha) + (1-\mu)(\beta-\beta_0)\psi_1(\beta) - (\tau-\alpha_0-\beta_0)\psi_1(\tau),$$

since $\alpha + \beta = \tau$.

Gradients of the ICRM Loss w.r.t. μ and τ Combining reconstruction and KL contributions:

$$\frac{\partial \mathcal{L}}{\partial \mu} = -\tau\,\psi_1(\mu\tau) \;+\; \lambda\,\tau[(\alpha-\alpha_0)\psi_1(\alpha) - (\beta-\beta_0)\psi_1(\beta)]\,, \tag{14a}$$

$$\frac{\partial \mathcal{L}}{\partial \tau} = -\mu\,\psi_1(\mu\tau) + \psi_1(\tau) \;+\; \lambda[\mu(\alpha-\alpha_0)\psi_1(\alpha) + (1-\mu)(\beta-\beta_0)\psi_1(\beta) - (\tau-\alpha_0-\beta_0)\psi_1(\tau)]\,. \tag{14b}$$

G Proof of Lemma 3.1

Proof. Recall (14a) and (14b) with $\alpha = \mu\tau$, $\beta = (1-\mu)\tau$. Define the tetragamma as $\psi_2(x) = d\psi_1(x)/dx$. As $\varepsilon = 1-\mu \to 0$, regularity at $\alpha \to \tau > 0$ gives

$$\psi_1(\mu\tau) = \psi_1(\tau) - \varepsilon\,\tau\,\psi_2(\tau) + O(\varepsilon^2) = \psi_1(\tau) + O(\varepsilon),$$

and the small-argument behavior at $\beta = \varepsilon\tau$ gives

$$\psi_1(\beta) = \psi_1(\varepsilon\tau) = \frac{1}{(\varepsilon\tau)^2} + O(1).$$

Hence

$$\tau[(\alpha-\alpha_0)\psi_1(\alpha) - (\beta-\beta_0)\psi_1(\beta)] = \frac{\beta_0}{\tau\,\varepsilon^2} \;-\; \frac{1}{\varepsilon} \;+\; O(1),$$

and

$$\frac{\partial \mathcal{L}}{\partial \tau} = \; O(\varepsilon) \;+\; \lambda\Big(-\frac{\beta_0}{\varepsilon\,\tau^2} + O(1)\Big)\,.$$

Finally, $\nabla_\theta \mu = \mu(1-\mu)\nabla_\theta \Delta u_\theta = (\varepsilon - \varepsilon^2)\,\nabla_\theta \Delta u_\theta$. Multiplying out gives

$$\frac{\partial \mathcal{L}}{\partial \mu}\,\nabla_\theta \mu = \Big(\frac{\lambda\beta_0}{\tau\varepsilon^2} - \frac{\lambda}{\varepsilon} - \tau\psi_1(\tau) + O(1)\Big)\cdot(\varepsilon-\varepsilon^2)(\nabla_\theta \Delta u_\theta) = \Big(\frac{\lambda\beta_0}{\tau\varepsilon} + O(1)\Big)\nabla_\theta \Delta u_\theta,$$

$$\frac{\partial \mathcal{L}}{\partial \tau}\,\nabla_\theta \tau = \Big(-\frac{\lambda\beta_0}{\varepsilon\tau^2} + O(1)\Big)\nabla_\theta \tau,$$

which yields the claim. □

H Proof of Theorem 3.2

Proof. Finiteness at an interior point and continuity. Let $\mu_0 = \alpha_0/(\alpha_0+\beta_0)$ and $\tau_0 = \alpha_0 + \beta_0$, so $(\alpha,\beta) = (\alpha_0,\beta_0)$ at (μ_0,τ_0). Then $\mathrm{KL}(\mathrm{Beta}(\alpha_0,\beta_0)\|\mathrm{Beta}(\alpha_0,\beta_0)) = 0$ and $-[\psi(\alpha_0) - \psi(\tau_0)] < \infty$, hence $\mathcal{L}(\mu_0,\tau_0) < \infty$. Because $(\mu,\tau) \mapsto (\alpha,\beta)$ is continuous on $(0,1)\times(0,\infty)$ and both ψ and the KL closed form are continuous on $(0,\infty)$, $\mathcal{L}$ is continuous.

Asymptotic tools. As $x \to 0^+$, $\psi(x) = -x^{-1} - \gamma + O(x)$ with γ as the Euler's constant; as $z \to \infty$, $\psi(z) = \log z - \frac{1}{2z} + O(z^{-2})$. Recall (8)

$$\begin{aligned}\mathrm{KL}\big(\mathrm{Beta}(\alpha,\beta) \,\|\, \mathrm{Beta}(\alpha_0,\beta_0)\big) &= \log \frac{\Gamma(\tau)}{\Gamma(\alpha)\Gamma(\beta)} - \log \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \\ &\quad + (\alpha-\alpha_0)\big[\psi(\alpha)-\psi(\tau)\big] + (\beta-\beta_0)\big[\psi(\beta)-\psi(\tau)\big].\end{aligned} \tag{15}$$

When $\tau \to \infty$ with $\mu = \alpha/\tau \in [\delta, 1-\delta] \subset (0,1)$,

$$\log \frac{\Gamma(\tau)}{\Gamma(\alpha)\Gamma(\beta)} = \alpha \log \frac{\tau}{\alpha} + \beta \log \frac{\tau}{\beta} + \tfrac{1}{2} \log \frac{\alpha\beta}{\tau} + O(1), \tag{16}$$

with $O(1)$ uniform in $\mu \in [\delta, 1-\delta]$.

Boundary coercivity. Let $(\mu_n, \tau_n) \in (0,1) \times (0,\infty)$ approach the boundary of $[0,1] \times [0,\infty]$. Passing to a subsequence, exactly one of the following disjoint regimes occurs:

(A) $\tau_n \to 0^+$; (B) $\tau_n \to \infty$; (C) $0 < \inf_n \tau_n \le \sup_n \tau_n < \infty$ and $\mu_n \to 0$ or 1.

Write $\alpha_n = \mu_n \tau_n$ and $\beta_n = (1-\mu_n)\tau_n$.

Case (A): $\tau_n \to 0^+$**.**

- If $\mu_n \to \mu \in (0,1)$, then $\alpha_n, \beta_n \to 0^+$ and

$$\psi(\tau_n) - \psi(\alpha_n) = \Big(-\tfrac{1}{\tau_n} + O(1)\Big) - \Big(-\tfrac{1}{\alpha_n} + O(1)\Big) = \frac{1-\mu}{\mu}\,\frac{1}{\tau_n} + O(1) \;\to\; \infty,$$

  so the $-[\psi(\alpha) - \psi(\tau)]$ term alone yields $\mathcal{L}(\mu_n, \tau_n) \to \infty$.

- If $\mu_n \to 0$, then $\alpha_n \to 0$ and

$$\psi(\tau_n) - \psi(\alpha_n) = \frac{1-\mu_n}{\mu_n}\,\frac{1}{\tau_n} + O(1) = \frac{1-\mu_n}{\alpha_n} + O(1) \;\to\; \infty,$$

  hence $\mathcal{L}(\mu_n, \tau_n) \to \infty$.

- If $\mu_n \to 1$, then $\beta_n \to 0$ and, from (15),

$$(\beta_n - \beta_0)\big[\psi(\beta_n) - \psi(\tau_n)\big] = -\beta_0\big[\psi(\beta_n) - \psi(\tau_n)\big] = \beta_0\left(\frac{1}{\beta_n} - \frac{1}{\tau_n} + O(1)\right) \;\to\; \infty,$$

  so again $\mathcal{L}(\mu_n, \tau_n) \to \infty$.

Case (B): $\tau_n \to \infty$**.**

(B1) If $\mu_n \in [\delta, 1-\delta]$ eventually for some $\delta \in (0, \frac{1}{2})$, then $\alpha_n, \beta_n \asymp \tau_n$. Insert (16) and the large–z digamma expansion into (15); all $O(\tau_n)$ terms cancel and, uniformly in $\mu_n \in [\delta, 1-\delta]$,

$$\mathrm{KL}(\mathrm{Beta}(\alpha_n,\beta_n) \,\|\, \mathrm{Beta}(\alpha_0,\beta_0)) = \tfrac{1}{2}\log \tau_n + O(1) \;\to\; \infty.$$

Meanwhile $\psi(\tau_n) - \psi(\alpha_n) = \log \tau_n - \log(\mu_n \tau_n) + O(1) = -\log \mu_n + O(1)$ is bounded on $[\delta, 1-\delta]$. Hence $\mathcal{L}(\mu_n, \tau_n) \to \infty$.

(B2) If $\mu_n \to 0$ (the case $\mu_n \to 1$ is symmetric), write $\alpha_n = \mu_n \tau_n$ and $\beta_n = \tau_n - \alpha_n$.

– If $\alpha_n \to a \in (0,\infty)$, expand only the large arguments τ_n, β_n in (15):

$$\log \frac{\Gamma(\tau_n)}{\Gamma(\beta_n)} = \alpha_n \log \beta_n + O(1) = \alpha_n \log \tau_n + O(1), \qquad \psi(\beta_n) - \psi(\tau_n) = O(\tau_n^{-1}),$$

and $(\alpha_n - \alpha_0)\big[\psi(\alpha_n) - \psi(\tau_n)\big] = -(\alpha_n - \alpha_0)\log \tau_n + O(1)$. Thus

$$\mathrm{KL}(\mathrm{Beta}(\alpha_n,\beta_n) \,\|\, \mathrm{Beta}(\alpha_0,\beta_0)) = \alpha_0 \log \tau_n + O(1) \;\to\; \infty,$$

so $\mathcal{L}(\mu_n, \tau_n) \to \infty$.

- If $\alpha_n \to 0$, then

$$(\alpha_n - \alpha_0)\big[\psi(\alpha_n) - \psi(\tau_n)\big] = -\alpha_0\big[\psi(\alpha_n) - \psi(\tau_n)\big] = \alpha_0\Big(\tfrac{1}{\alpha_n} + \log\tau_n + O(1)\Big) \;\to\; \infty,$$

  hence $\mathrm{KL} \to \infty$ and $\mathcal{L} \to \infty$.
- If $\alpha_n \to \infty$ while $\mu_n = \alpha_n/\tau_n \to 0$, then

$$\psi(\tau_n) - \psi(\alpha_n) = \log\tau_n - \log\alpha_n + o(1) = -\log\mu_n + o(1) \;\to\; \infty,$$

  so $\mathcal{L}(\mu_n, \tau_n) \to \infty$.

Case (C): $0 < \inf_n \tau_n \le \sup_n \tau_n < \infty$ and $\mu_n \to 0$ or 1. By symmetry, take $\mu_n \to 0$. Then $\alpha_n = \mu_n\tau_n \to 0$ while $\psi(\tau_n) = O(1)$, hence

$$\psi(\tau_n) - \psi(\alpha_n) = O(1) - \Big(-\tfrac{1}{\alpha_n} + O(1)\Big) = \tfrac{1}{\alpha_n} + O(1) \;\to\; \infty,$$

and therefore $\mathcal{L}(\mu_n, \tau_n) \to \infty$.

Compact sublevel sets and attainment. From the three regimes, any sequence with $\mathcal{L}(\mu_n, \tau_n) \le c$ stays a positive distance from $\{\mu = 0, 1\} \cup \{\tau = 0\}$ and also has $\sup_n \tau_n < \infty$. Hence $\{\mathcal{L} \le c\} \subset [\varepsilon, 1-\varepsilon] \times [\varepsilon, M]$ for some $\varepsilon, M > 0$, a compact rectangle contained in $(0,1) \times (0,\infty)$. By continuity (Weierstrass), $\mathcal{L}$ attains its minimum there; consequently any minimizer lies in the open domain $(0,1) \times (0,\infty)$. $\square$