IN-CONTEXT BAYESIAN REWARD MODELING FOR TEST-TIME STEERABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward models (RMs) are central to aligning language models via reinforcement learning (RL), yet conventional classifier RMs are static after training and tied to their training data distribution. This limits generalization to unseen preference types, which is an increasingly salient need with the emergence of verifiable rewards. To address this gap, we introduce **Variational In-Context Reward Modeling (ICRM)**, which casts reward modeling as amortized variational inference over a latent preference probability conditioned on few-shot, in-context preference demonstrations, with a conjugate Beta prior on the Bradley–Terry model. ICRM employs a two-headed regressor that decouples a preference mean (μ) from a confidence factor (τ), jointly parameterizing a Beta posterior given demonstrations and enabling **test-time steerability of RMs**. On reward model benchmarks, a *fixed* ICRM improves accuracy simply by increasing demonstrations from 1 to 8, achieving $63.7\% \rightarrow 95.6\%$ on the Focus subset in RewardBench 2, for instance. In reinforcement learning with verifiable reward (RLVR) experiment for math reasoning, ICRM used as the reward yields faster and higher accuracy improvements than a verifier-based reward, reaching stronger performance with 50% of training problems compared to the verifier-based reward. Finally, we provide theoretical guarantees that the global minimizer of our loss admits finite confidence and show analytically how KL regularization tempers over-optimization. Together, ICRM offers a practical and principled framework for a test-time steerable reward model that generalizes beyond the training distribution.

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 & 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 & 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 during 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., 2024b) offers flexible evaluation criteria with strong performance (Lambert et al., 2025; Malik et al., 2025; Liu et al., 2025b), these gains often rely on large proprietary models such as Gemini 2.5 (Comanici et al., 2025) and GPT-4o (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.

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 Bradley–Terry model yields a closed-form training loss via variational inference. This variational loss enables ICRM to learn preferences *in-context* with few-shot demonstrations,

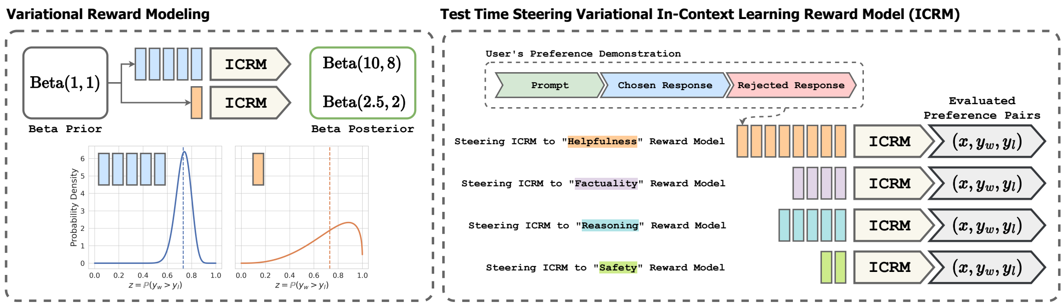


Figure 1: Variational in-context reward modeling (**ICRM**) with Beta prior for the Bradley-Terry (BT) model. ICRM directly models the mean and sharpness of the Beta posterior, calibrated to how "*confident*" the model is for the preference triplet (x, y_w, y_l) given in-context preference demonstrations. This yields **test-time steerability of the reward model** for *any* preferences or tasks.

allowing **test-time steerability of a classifier RM** that can dynamically adapt to arbitrary preferences, *e.g.,* reasoning accuracy or factuality. Furthermore, we prove that a KL penalty to the Beta prior tempers the learned preference mean and yields a global interior optimum. Comprehensively, our main contributions are summarized below:

1. **Test-time steerability of ICRM to arbitrary preferences** (Sections 4.3 and 4.4): On RM-Bench (Liu et al., 2025b) and RewardBench 2 (Malik et al., 2025), a fixed ICRM shows consistent gains from one to eight in-context preference demonstrations (N) across distinct domains, *e.g.,* $55.5\% \rightarrow 98.3\%$ for "Precise IF" and $79.3\% \rightarrow 94.9\%$ in average with increasing N.
2. **Versatility of ICRM in RL, including verifiable rewards** (Section 5): Using eight demonstrations of accurate and inaccurate reasoning trajectories as preference pairs, ICRM's reward scores calibrate to the accuracy, leading to higher reasoning performance of the policy, achieved with 50% of prompts compared to Reinforcement Learning with Verifiable Reward (RLVR).
3. **Theoretical mitigation of over-optimization via KL regularization** (Section 6): We prove that regularizing the Beta posterior by a uniform Beta prior guarantees a global interior optimum, thereby tempering excessive maximization of the preference mean on training data.

2 Background

2.1 Preliminaries

A classifier reward model (RM), $r_\theta(x, y)$, is a function parameterized by θ that outputs a scalar score indicating the quality of a response y given a prompt x (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 & 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:

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

Once the preference distribution shown in the training set $\mathcal{D}_{\text{train}}$ is encoded into θ via fine-tuning, it cannot be adaptively updated at test time without additional retraining, significantly limiting the flexibility of classifier RMs.

2.2 Theoretical Background

In-context learning as implicit fine-tuning Recent works show that in-context learning (ICL) in large language models (LLMs) adapts them to new texts with few-shot examples, similar to how models learn through 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.

Estimating the true preference distribution in the Bradley–Terry model 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 Bradley–Terry (BT) setting, the maximum-likelihood estimator (MLE) exists and enjoys consistency and asymptotic normality. Consequently, 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 further study the practical parameterization of the BT model and its application in Appendix B.

Bayesian treatment of the Bradley-Terry model A Bayesian treatment of the BT model necessitates the selection of a suitable prior distribution for its parameters (Chen & Smith, 1984; Whelan, 2017; Wainer, 2023; Fageot et al., 2024). The general form of the model, which accommodates $N > 2$ 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 Dirichlet prior (Chen & Smith, 1984).

3 Variational In-Context Reward Modeling

To address the limitations of classifier RMs, 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 approximating the true preference distribution with a Beta posterior conditioned on in-context preference demonstrations and placing a Beta prior for gradual regularization.

3.1 Problem Setup

Prior distribution 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., $z := P(y_w \succ y_l \mid x, \mathcal{C}) \in [0, 1]$. 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, $p(z \mid x, y_w, y_l, \mathcal{C})$, reflecting 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 distribution By Bayes' rule, the *true posterior* over z after observing o is:

$$p(z \mid o, x, y_w, y_l, \mathcal{C}) \propto p(o \mid z, x, y_w, y_l, \mathcal{C}) \cdot p(z \mid x, y_w, y_l, \mathcal{C}). \tag{4}$$

which is our inferential target. However, computing 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 $o = \mathbb{1}_{y_w \succ y_l}$. Therefore, 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 In-Context Reward Modeling as Variational Inference

We parameterize $q_\theta(z \mid x, y_w, y_l, \mathcal{C})$ with the autoregressive language model θ (Radford et al., 2019), which directly maps the inputs to the parameters of an approximate posterior distribution, namely the in-context reward modeling (ICRM). Specifically, they are designed as a classifier RM in equation 1. In this section, we outline the choice of the prior distribution and propose the final learning objective for ICRM as variational inference.

Beta prior for the Bradley-Terry model Extending from the discussion on the Bayesian treatment of the BT model, we propose the use of a Beta prior in 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., 2024a; 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 probability 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 utility scores $u(x, y_w, \mathcal{C})$ and $u(x, y_l, \mathcal{C})$ and confidence scores $s(x, y_w, \mathcal{C})$ and $s(x, y_l, \mathcal{C})$, shortened as u_w, u_l, s_w, and s_l. We reparameterize the Beta posterior parameters as:

$$\begin{cases} \alpha_q = \mu\tau, \\ \beta_q = (1-\mu)\tau \end{cases} \quad \text{where} \quad \begin{cases} \mu = \sigma(u_w - u_l), \\ \tau = \mathrm{Softplus}(s_w) + \mathrm{Softplus}(s_l) + 1, \end{cases} \tag{5}$$

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 Bradley–Terry 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 provided by the demonstrations.

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:

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

The first term in equation 6, $-\mathbb{E}_{q_\theta(z|x,y_w,y_l,\mathcal{C})}[\log z]$, 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; Kim et al., 2024a). Meantime, τ controls how sharply the distribution concentrates around this preference.

The second term in equation 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 model's output, preventing the posterior from deviating excessively from the prior, especially when little contextual evidence is available (Joo et al., 2020), *e.g.,* N is small. The KL-divergence between two Beta distributions, $p = \mathrm{Beta}(\alpha_p, \beta_p)$ and $q = \mathrm{Beta}(\alpha_q, \beta_q)$, also has a closed-form solution (Loaiza-Ganem & Cunningham, 2019; Joo et al., 2020):

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

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:

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

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 equation 9, the divergence penalty can be controlled with the pre-defined prior distribution $p = \text{Beta}(\alpha_0, \beta_0)$. If we have explicitly collected annotations for the pair (x, y_w, y_l) for given few-shot examples $\mathcal{C}$, we may set different (α_0, β_0) per item. However, it is typically hard to collect such data. For this reason, 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 or voting over multiple preference models can be used to generate such data to provide a more informative prior (Yang et al., 2024c; Singh et al., 2025).

4 Experiments

We validate the variational in-context reward models (ICRM) from two perspectives. Given a *single* trained ICRM, we analyze if they can dynamically adapt to users' preferences *on the fly*:

1. **Test-Time Steerability**: Does the posterior mean $\mathbb{E}_{q_\theta}[z] = \mu$ adapt to the implicit preference distribution induced by in-context demonstrations $\mathcal{C}$?
2. **Confidence Calibration**: Does the posterior concentration τ increase appropriately with $|\mathcal{C}|$, yielding sharper Beta posteriors as more demonstrations are provided?

We visit the first research question by testing if ICRM can achieve gradually better performance in reward model benchmarks with an increasing number of samples. Meanwhile, we trace the confidence factor from the model, analyzing the confidence calibration.

4.1 Training Setup

Model We experiment with two base model families, Qwen3-4B-Base (Yang et al., 2025) and Llama-3.2-3B-Base (Dubey et al., 2024). To control prior preference distributions, we train on top of the pre-trained checkpoints. The projection head $W_p \in \mathbb{R}^{d_{\text{model}} \times 2}$ is initialized with $\mathcal{N}(0, (d_{\text{model}} + 1)^{-1})$ following Stiennon et al. (2020); Huang et al. (2024); Hong et al. (2025).

Training data Reward models (RMs) are trained on Skywork-Preferences-v0.2 (Liu et al., 2024), a mixture of MagPie (Xu et al., 2025), WildGuard (Han et al., 2024), OffsetBias (Park et al., 2024), and HelpSteer 2 (Wang et al., 2025a), covering diverse domains of human preference learning. We assume each dataset reflects a consistent implicit preference distribution, *e.g.,* WildGuard has a consistent preference bar for safety. 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 disjointly from the training row used for learning. To reduce template bias, we adopt a minimal prompt format without explicit instructions in Appendix C. Additional details for training configurations are listed in Appendix D.

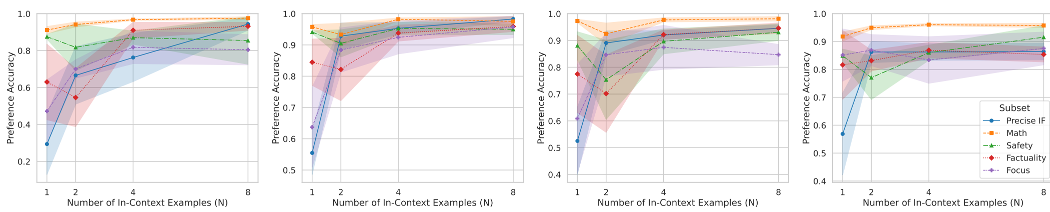
4.2 Evaluation Setup

Given a *single* model trained as ICRM, we evaluate the test-time steerability across different domains by providing domain-specific in-context demonstrations $\mathcal{C}$. Here, $\mathcal{C}$ and the tested preference pair should share the *same* preference distribution to evaluate the in-context preference learning.

Evaluation data We evaluate using RewardBench 2 (Malik et al., 2025) and RM-Bench (Liu et al., 2025b), which spans six domains of preference learning and covers varying difficulty, respectively. Each subset is treated as a coherent preference domain, and ICRM adapts at test-time via few-shot in-context learning only. For baselines, we fix the training dataset to Skywork-Preferences-v0.2 (Liu

	RM-Bench				RewardBench 2						
	Easy	Normal	Hard	Avg.	Fact.	Prec. IF	Math	Safety	Focus	Ties	Avg.
Bradley-Terry (Liu et al., 2024)	**89.3**	**75.8**	52.6	$\underline{70.2}$	69.7	40.6	60.1	94.2	94.1	71.7	71.8
GRM (Yang et al., 2024d)	$\underline{86.2}$	70.6	45.1	67.3	62.7	35.0	58.5	92.2	89.3	68.2	67.7
URM (Lou et al., 2025)	84.0	73.2	53.0	70.0	68.8	45.0	63.9	91.8	**97.6**	76.5	73.9
ICRM (Llama-3.2-3B)											
$N = 1$	$61.0_{0.05}$	$62.9_{0.01}$	$53.5_{0.05}$	59.1	$69.1_{0.11}$	$46.5_{0.24}$	$92.3_{0.04}$	$42.8_{0.19}$	$63.3_{0.16}$	$63.8_{0.15}$	63.0
$N = 2$	$63.5_{0.08}$	$65.2_{0.01}$	$62.1_{0.03}$	63.6	$87.1_{0.10}$	$66.0_{0.33}$	$95.6_{0.02}$	$70.9_{0.05}$	$77.2_{0.19}$	$60.7_{0.08}$	76.3
$N = 4$	$66.0_{0.02}$	$66.3_{0.02}$	$\mathbf{66.9}_{0.04}$	66.4	$87.6_{0.27}$	$86.3_{0.07}$	$93.7_{0.02}$	$68.9_{0.16}$	$75.4_{0.13}$	$70.6_{0.08}$	80.4
$N = 8$	$72.1_{0.05}$	$66.0_{0.01}$	$\underline{66.4}_{0.07}$	68.2	$88.6_{0.03}$	$90.3_{0.05}$	$96.2_{0.04}$	$89.9_{0.07}$	$\underline{96.4}_{0.01}$	$70.9_{0.04}$	88.7
ICRM (Qwen3-4B)											
$N = 1$	$71.5_{0.05}$	$66.3_{0.01}$	$58.8_{0.05}$	65.5	$84.5_{0.10}$	$55.5_{0.10}$	$95.8_{0.01}$	$94.1_{0.01}$	$63.7_{0.17}$	$82.3_{0.05}$	79.3
$N = 2$	$70.0_{0.14}$	$67.3_{0.06}$	$62.8_{0.06}$	66.7	$82.2_{0.13}$	$92.8_{0.06}$	$93.3_{0.05}$	$90.5_{0.06}$	$88.4_{0.09}$	$83.1_{0.04}$	88.4
$N = 4$	$75.5_{0.08}$	$71.2_{0.02}$	$62.9_{0.07}$	69.9	$\underline{93.8}_{0.03}$	$\underline{95.3}_{0.02}$	$\mathbf{98.2}_{0.01}$	$\mathbf{95.3}_{0.03}$	$92.4_{0.07}$	$\underline{83.7}_{0.04}$	$\underline{93.1}$
$N = 8$	$84.3_{0.02}$	$\underline{73.4}_{0.01}$	$64.6_{0.07}$	**74.0**	$\mathbf{95.9}_{0.02}$	$\mathbf{98.3}_{0.01}$	$\underline{97.5}_{0.02}$	$\underline{95.0}_{0.02}$	$95.6_{0.05}$	$\mathbf{87.2}_{0.02}$	**94.9**

Table 1: Reward model (RM) benchmark evaluation for RMs trained on Skywork-Preference-v0.2. We evaluate ICRM with a 5-fold, using an isolated fold as the demonstration pool and evaluating the remaining folds. The highest and the second-highest are **bold** and underlined by column.



(a) Bradley-Terry (b) Variational ($\lambda = 0.1$) (c) Variational ($\lambda = 0.5$) (d) Variational ($\lambda = 1.0$)

Figure 2: **Preference Accuracy** – Subset-level ablation study on λ of ICRM on RewardBench 2.

et al., 2024), selecting three reward models trained with different objectives: (1) plain Bradley-Terry model, (2) GRM (Yang et al., 2024d), and (3) URM (Lou et al., 2024).

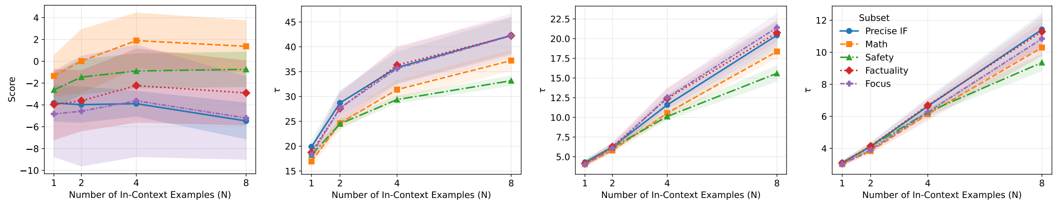
K-fold evaluation We partition each subset into 5 folds. In each run, one fold serves *only* as the held-out demonstration set and the remaining four folds are used for evaluation. We randomly sample preference pairs (x, y_w, y_l) from the demonstration set with the size of $N \in \{1, 2, 4, 8\}$ to curate the final in-context demonstration prefix $\mathcal{C}$, comparing the reward scores obtained from $(x', y'_w, \mathcal{C})$ and $(x', y'_l, \mathcal{C})$ where (x', y'_w, y'_l) is sampled from the four folds. After iterating over five folds, the final results are reported as the mean and standard deviation across the 5 folds per N.

4.3 Implicit Preference Steerability

ICRM improves preference accuracy with increasing in-context demonstrations Table 1 presents evaluation results for ICRM trained on Llama-3.2-3B and Qwen3-4B base models with $\lambda = 0.1$. For both models, accuracy increases monotonically as the number of demonstrations N grows. Through difficulty-based comparison on RM-Bench, ICRM demonstrates gradual improvements across levels, eventually surpassing the baselines with Qwen3-4B ICRM and $N = 8$ by 4%.

In the meantime, domain-wise comparison through RewardBench 2 supports test time adaptation of ICRM. On the “Precise IF” subset, which evaluates compliance with complex instruction constraints, accuracy starts close to random guessing (46.5% for Llama-3.2 and 55.5% for Qwen3) but reaches 98% with $N = 8$. This illustrates the Beta posterior being progressively specified with more evidence from an initially uniform prior. On average, it is notable that in-distribution preference demonstrations can boost the performance of ICRM up to 94.9% from 79.3% with $N = 8$, surpassing the first-ranked Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025a) that achieves 84.1%.

Ablation on λ Figure 2 shows the effect of λ on in-context learning, including the plain Bradley–Terry (BT) baseline. In Figure 2a, we test if the BT can also learn the preference in-context when trained on our custom preference data in Section 4.1. Interestingly, BT also benefits from in-context learning under our formulation, improving steadily with N in Figure 2a.



(a) Bradley-Terry (b) Variational ($\lambda = 0.1$) (c) Variational ($\lambda = 0.5$) (d) Variational ($\lambda = 1.0$)

Figure 3: **Calibration** – Confidence factor τ of the posterior $\text{Beta}(\alpha_q, \beta_q)$ with varying λ and N.

Comparing Figures 2b to 2d, the variance of preference accuracy increases with larger λ. For example, $\lambda = 0.1$ converges near 95% with progressively tighter error bars, whereas $\lambda = 0.5$ and $\lambda = 1.0$ exhibit wide error ranges even at $N = 8$. This reflects the role of the KL penalty term: stronger regularization pulls ICRM closer to the uninformative prior $\text{Beta}(1, 1)$. Thus, $\lambda = 0.1$ yields the best balance between uncertainty at $N = 1$ and confident convergence with more demonstrations.

4.4 Confidence Calibration

Confidence scales with the number of demonstrations Figure 3 shows that ICRM's confidence factor τ increases with the number of in-context demonstrations N. Similar to the ablation study in Section 4.3, we report the score distribution of the BT-based ICRM in Figure 3a, which is insensitive to N and thus uncalibrated. The monotone rise of τ with increasing N from Figures 3b to 3d is consistent across subsets, indicating that the model not only adjusts its mean preference but also encodes higher certainty as more evidence accumulates. In effect, τ serves as a task-agnostic proxy for contextual strength, thereby providing a calibrated measure of model confidence at test-time.

Stronger KL penalty lowers test-time confidence In Figure 3, the magnitude of τ decreases as λ increases. For instance, τ grows up to 45 in Figure 3b, while it reaches a maximum of 12 in Figure 3d. Recall that the variance of the approximated Beta posterior $z \sim \text{Beta}(\alpha_q, \beta_q)$ is

$$\text{Var}_{q_\theta}[z] = \frac{\alpha_q \beta_q}{(\alpha_q + \beta_q)^2(\alpha_q + \beta_q + 1)} = \frac{\mu(1-\mu)}{\tau + 1} \quad \text{where} \quad \alpha_q = \mu\tau \quad \text{and} \quad \beta_q = (1-\mu)\tau, \tag{10}$$

decreases monotonically with τ; hence larger τ encodes sharper, more confident beliefs around μ, while a stronger KL (larger λ) suppresses τ and reduces confidence when context is scarce.

5 In-Context Reward Model in Reinforcement Learning

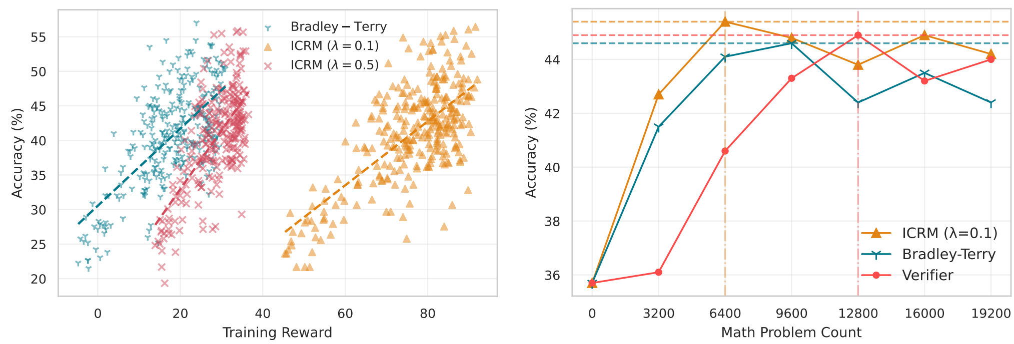
Motivated by the steerability results in Section 4.3, we investigate whether ICRM can parameterize *arbitrary preferences* via few-shot demonstrations in a full RLHF setting in Figure 4. Our variational construction naturally provides a principled extension of scoring in-context reward modeling. The approximate posterior $q_\theta(z \mid x, y_w, y_l, \mathcal{C}) = \text{Beta}(\alpha_q, \beta_q)$ is parameterized by equation 5, where μ encodes the expected preference and τ the confidence via response-specific scores $u_\theta(x, y, \mathcal{C})$ and $s_\theta(x, y, \mathcal{C})$. For a single (x, y), we interpret u_θ as the local contribution to μ and s_θ as the contribution to τ, and define the stand-alone reward:

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

Intuitively, $R_{\text{ICRM}}(x, y, \mathcal{C})$ 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.

5.1 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 Qwen2.5-1.5B-Base (Qwen et al., 2025)



(a) Reward to Accuracy

(b) Evaluation Accuracy

Figure 4: **Parameterizing Verifiable Reward with ICRM** - Reinforcement learning (RL) training on math reasoning with ICRM, Bradley-Terry reward, and verifier-based reward. Correlation between reward and accuracy (Figure 4a) and evaluation accuracy during the training (Figure 4b).

on INTELLECT-MATH[1] using GRPO (Shao et al., 2024) under three reward configurations: (1) **ICRM**: Qwen3-4B-Base as the ICRM RM with $\lambda = 0.1$ and $N = 8$ demonstrations drawn from the "Math" subset of RewardBench 2; (2) **Bradley-Terry reward**: Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024) trained on the same preference data; and (3) **Verifier-based reward**: exact-match supervision against gold answers. Additional training details are provided in Appendix E.

5.2 Results

ICRM's reward scores are aligned with gold accuracy in RLVR In Figure 4, 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. Figure 4a and Table 2 analyze 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 $\lambda = 0.1$ generally has a stronger alignment with the gold accuracy. Thus, the results indicate that even the deterministic accuracy in reasoning tasks can be modeled via ICRM.

	Pearson r	R^2_{OLS}	R^2_{Iso}
ICRM ($\lambda = 0.1$)	**0.691**	**0.477**	0.459
ICRM ($\lambda = 0.5$)	0.685	0.469	**0.461**
Bradley-Terry	0.663	0.439	0.428

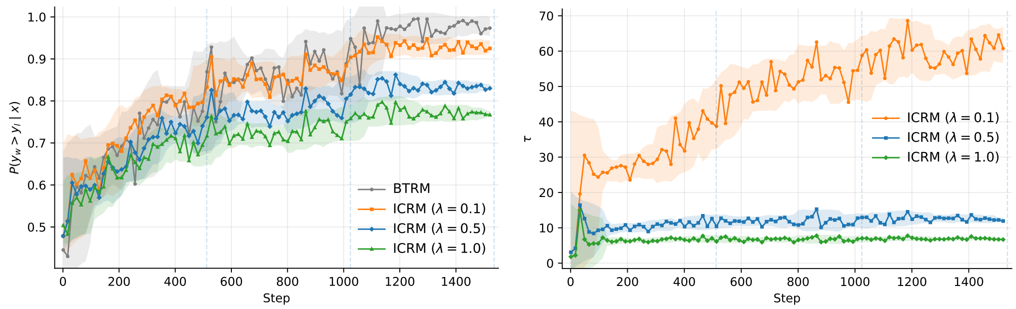
Table 2: Statistical analysis for the alignment between reward model scores and verified accuracy.

Faster accuracy convergence with ICRM than the gold verifier as a reward In Figure 4b, we track the policies trained with each reward on MATH-500 (Lightman et al., 2024) every 50 gradient updates. We report the average scores of five rollouts. Notably, the policy trained with ICRM demonstrated the stiffest accuracy increase in the initial training, compared to those of Bradley-Terry (BTRM) and verifier-based reward. With ICRM, the policy achieved an accuracy of up to 45.4% on the 100th step, whereas it was at most 45.0% and 44.6% for verifier and BTRM cases, respectively. Overall, by achieving the best evaluation accuracy with the least training data, ICRM has a practical advantage in effectively modeling arbitrary preferences simply with a few-shot demonstrations.

6 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

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



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

Figure 5: The learning curve of the learned preference mean μ (Figure 5a), the concentration factor τ of the parameterized Beta posterior (Figure 5b) in the variational in-context reward modeling.

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 6.1 (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 $\varepsilon \to 0^+$,*

$$\nabla_\theta \mathcal{L}_{\text{ICRM}} = \underbrace{\left(\frac{\lambda\beta_0}{\varepsilon\tau} + O(1)\right)}_{\text{Utility Coefficient}} \nabla_\theta \Delta u_\theta \; + \; \underbrace{\left(-\frac{\lambda\beta_0}{\varepsilon\tau^2} + O(1)\right)}_{\text{Confidence Coefficient}} \nabla_\theta \tau.$$

Δu_θ, the learned preference margin, 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. We provide the proof in Appendices F and G. Since the Lemma 6.1 is stated for finite τ, we next prove that the global minimizer indeed has $0 < \tau^\star < \infty$.

Theorem 6.2. *Assume $\lambda > 0$ and $\alpha_0, \beta_0 > 0$. For $(\mu, \tau) \in (0, 1) \times (0, \infty)$, every global minimizer $(\mu^\star, \tau^\star)$ of $\mathcal{L}_{ICRM}(\mu, \tau; \alpha_0, \beta_0)$ defined in equation 9 satisfies*

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

Consequently, the optimizer cannot place mass at the preference edges ($\mu \in \{0, 1\}$), nor can it collapse or diverge in confidence ($\tau \in \{0, \infty\}$), thereby providing a theoretically guaranteed prevention of reward model over-optimization via preference mean tempering. See Appendix H for proof.

KL penalty provides controllable tempering of preference mean In Figure 5, we conduct an ablation study over $\lambda \in \{0.1, 0.5, 1.0\}$ along with the plain BT. With a larger λ, the convergence point of $P_\theta(y_w \succ y_l|x)$ in Figure 5a is smaller, demonstrating tempered preference means with stronger regularization. Furthermore, the confidence factor τ monotonically increases with weaker regularization, *i.e.,* smaller λ, allowing context-dependent calibration instead of divergence to $\tau \to \infty$. Overall, the training dynamics in Figure 5 align with the implications of the theoretical analysis: the regularization term tempers over-confidence for the training dataset with a global interior optimum.

7 Conclusion

In this work, we introduced **Variational In-Context Reward Modeling (ICRM)**, a Bayesian reward modeling scheme that yields the test-time steerability of classifier RM by viewing Bradley–Terry (BT) preferences as a latent probability with a Beta posterior conditioned on few-shot preference demonstrations. A controllable KL regularizer to a uniform Beta prior calibrates confidence and theoretically mitigates over-optimization, leading to gradual improvement with increasing number of demonstrations (N). We empirically validate the in-context preference learning ability via two reward model benchmarks, achieving 63.0% to 88.7% and 79.3% to 94.9% by simply increasing N from 1 to 8 in RewardBench 2. Furthermore, in reinforcement learning with verifiable rewards (RLVR) for math reasoning, ICRM parameterizes deterministic accuracy as preference with 8-shot preference demonstrations and accelerates accuracy gains relative to the verifier-based reward. Overall, ICRM is an effective, theoretically grounded reward model that adapts to *arbitrary preferences* once trained, from human preferences to verifiable rewards.

Reproducibility statement

We report the overall training details used in the paper, including the variational in-context reward model (ICRM) training and reinforcement learning with verifiable reward (RLVR) experiments. For ICRM training, along with the training template in Appendix C, we report the hardware details, model configurations, and optimizer settings in Appendix D with the code. For RLVR experiments, we report the hyperparameter settings and necessary code dependencies in Appendix E. Furthermore, we report the mean and standard deviation over five runs, including the five-fold evaluation, for the performance evaluations to ensure the reproducibility of the experiments.

References

Zachary Ankner, Mansheej Paul, Brandon Cui, Jonathan D. Chang, and Prithviraj Ammanabrolu. Critique-out-loud reward models. *arXiv preprint arXiv:2408.11791*, 2024. URL https://arxiv.org/abs/2408.11791.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Heejong Bong and Alessandro Rinaldo. Generalized results for the existence and consistency of the MLE in the bradley-terry-luce model. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2160–2177. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bong22a.html.

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. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

C Chen and Theodore 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.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, and Kornraphop Kawintiranon 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.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Tri 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.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak 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.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson,

K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. Learning without training: The implicit dynamics of in-context learning, 2025. URL https://arxiv.org/abs/2507.16003.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, and Isabel Kloumann et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alexander Nicholas D'Amour, Krishnamurthy Dj Dvijotham, Adam Fisch, Katherine A Heller, Stephen Robert Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan 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.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024.

Julien Fageot, Sadegh Farhadkhani, Lê-Nguyên Hoang, and Oscar 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.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Aman Gupta, Shao Tang, Qingquan Song, Sirou Zhu, Jiwoo Hong, Ankan Saha, Viral Gupta, Noah Lee, Eunki Kim, Siyu Zhu, Parag Agrawal, Natesh S. Pillai, and Sathiya Keerthi. AlphaPO: Reward shape matters for LLM alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=LmdZ0pSWtG.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=Ich4tv4202.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, W. Bradley Knox, and Dorsa 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.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings* *of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.626.

Jiwoo Hong, Noah Lee, Eunki Kim, Guijin Son, Woojin Chung, Aman Gupta, Shao Tang, and James 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.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. Liger kernel: Efficient triton kernels for llm training, 2024. URL https://arxiv.org/abs/2410.10989.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis 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.

Taejong Joo, Uijung Chung, and Min-Gwan Seo. Being Bayesian about categorical probability. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4950–4961. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/joo20a.html.

Kyuyoung Kim, Ah Jeong Seo, Hao Liu, Jinwoo Shin, and Kimin Lee. Margin matching preference optimization: Enhanced model alignment with granular feedback. In *The 2024 Conference on Empirical Methods in Natural Language Processing*, 2024a. URL https://openreview.net/forum?id=jmLKEtZsxN.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=8euJaTveKw.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 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.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.96/.

Andrew K. Lampinen, Arslan Chaudhry, Stephanie C. Y. Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James 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.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion 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.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-Reward: Bag of Tricks for Reward Modeling in LLMs, October 2024. URL http://arxiv.org/abs/2410.18451. arXiv:2410.18451 [cs].

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*, 2025a.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi 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.

Gabriel Loaiza-Ganem and John 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 (eds.), *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.

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

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

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*, 2024. URL https://arxiv.org/abs/2410.00847.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. Uncertainty-aware reward model: Teaching reward models to know what is unknown, 2025. URL https://arxiv.org/abs/2410.00847.

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

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation, 2025. URL https://arxiv.org/abs/2506.01937.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024. URL https://arxiv.org/abs/2402.08114.

Abhishek Naik, Yi Wan, Manan Tomar, and Richard S. Sutton. Reward centering. *Reinforcement Learning Journal*, 4:1995–2016, 2025.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron 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, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza

Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.

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.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: In-context learning of representations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=pXlmOmlHJZ`.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. OffsetBias: Leveraging debiased data for tuning evaluators. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1043–1067, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.57. URL `https://aclanthology.org/2024.findings-emnlp.57/`.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=HPuSIXJaa9`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL `https://arxiv.org/abs/2402.03300`.

Anikait Singh, Sheryl Hsu, Kyle Hsu, Eric Mitchell, Stefano Ermon, Tatsunori Hashimoto, Archit Sharma, and Chelsea 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`.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Hao Sun, Yunyi Shen, and Jean-Francois 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.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aKkAwZB6JV.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl, 2020.

Jacques 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.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii 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*, 2024a. URL https://openreview.net/forum?id=PvVKUFhaNy.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=MnfHxPP5gs.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages, 2025b. URL https://arxiv.org/abs/2505.11475.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alexander Nicholas D'Amour, Sanmi Koyejo, and Victor 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*, pp. 51161–51176. PMLR, July 2024b. URL https://proceedings.mlr.press/v235/wang24ay.html.

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

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pnk7vMbznK.

Adam X. Yang, Maxime Robeyns, Thomas Coste, Zhengyan Shi, Jun Wang, Haitham Bou-Ammar, and Laurence Aitchison. Bayesian reward models for llm alignment. *arXiv preprint arXiv:2402.13210*, 2024a. URL https://arxiv.org/abs/2402.13210.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence 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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Joshua C. Yang, Damian Dailisan, Marcin Korecki, Carina I. Hausladen, and Dirk 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`.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong 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`.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Boji Shan, Zeyuan Liu, Jia Deng, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. Advancing LLM reasoning generalists with preference trees. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=2ea5TNVR0c`.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, July 2024. URL `https://proceedings.mlr.press/v235/yuan24d.html`.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen 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`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Zhang, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. URL `https://arxiv.org/abs/2306.05685`.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *First Conference on Language Modeling*, 2024. URL `https://openreview.net/forum?id=GqDntYTTbk`.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.

A LIMITATIONS

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 16 in-context demonstrations, which is expected to result in a stronger performance based on the experimental results.

B RELATED WORKS

Preference data for reward modeling Reward models (RMs) in the reinforcement learning with human feedback (RLHF) pipeline act as human preference proxies, 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 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 a strong performance of reward models in practice.

Reward modeling in reinforcement learning with human feedback In parallel, previous works propose different 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., 2024b), 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) explores the generalized application of the BT model in language model reward modeling, such as comprising preference pairs across different prompts.

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., 2023). Robustness-oriented designs include energy-based RMs that refine scores via distributional modeling and conflict-aware filtering (Lochab & 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.

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.

C 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...

D 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 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 & 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 codes (von Werra et al., 2020). We use four NVIDIA H100 (80GiB) GPUs for ICRM training.

E Training Configurations for Reinforcement Learning

In general, we follow the optimizer and distributed training settings from Appendix D. 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[2] 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.

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

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

Table 3: Hyperparameters for GRPO training in Section 5.

F GRADIENT ANALYSIS OF ICRM LOSS

Recall equation 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

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

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

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.

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

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:

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

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

G PROOF OF LEMMA 6.1

Proof. Recall equation 14a and equation 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

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

and

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

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. $\square$

H PROOF OF THEOREM 6.2

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 equation 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:

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

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 equation 15,
$$(\beta_n - \beta_0)\big[\psi(\beta_n) - \psi(\tau_n)\big] = -\beta_0\big[\psi(\beta_n) - \psi(\tau_n)\big] = \beta_0\bigg(\frac{1}{\beta_n} - \frac{1}{\tau_n} + O(1)\bigg) \;\to\; \infty,$$
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 equation 16 and the large–z digamma expansion into equation 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 equation 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 $(0,1) \times (0,\infty)$. $\square$