Reward Model Routing in Alignment

Xinle Wu, Yao Lu
National University of Singapore

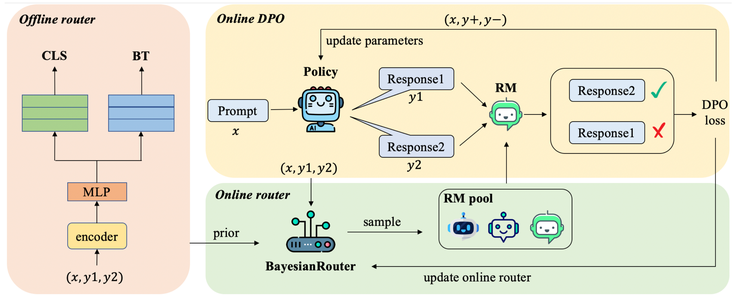
Motivation

(a) Limitations of Single Reward Models (RMs): No single RM consistently excels across all diverse tasks and domains (e.g., math vs. safety).High risk of reward hacking and overoptimization due to a single model's idiosyncratic biases.

(b) Inefficiency of Standard Ensembling: Running multiple RMs in parallel for every query incurs an expensive $O(N)$ inference cost.

(c) Flaws in Existing Routing Methods: Exploration: Current bandit algorithms (like LinUCB) may prematurely lock onto suboptimal RMs. Cold-Start Problem: Routers often start with no prior knowledge, leading to inefficient early-stage training.

BayesianRouter



Phase 1: Offline RM Strengths Learning

(a) Goal: Learn reward model (RM) specializations from static human-annotated datasets.

(b) Multi-Task Training: Bradley-Terry (to rank RM competence) and Classification (to predict RM correctness) .

Phase 2: Online Bayesian Selection

(a) Goal: Dynamically adapt RM selection to the evolving policy during alignment.

(b) Mechanism: Bayesian Thompson Sampling for per-query (instance-level) RM selection.

(c) Prior Injection: Bootstraps the online router using the offline embeddings as Gaussian priors.

Experiments & Performance

Table 1: Main results on instruction-following and reasoning benchmarks.

Method	Instruction-Following			Reasoning	
	AlpacaEval-2	MT-Bench	Chat-Arena-Hard	GSM8K	MMLU
SFT	50.00	50.00	50.00	67.63	54.29
RM0	56.02	52.50	59.60	72.78	56.00
RM1	61.86	56.25	64.80	74.22	57.03
RM2	59.50	53.75	63.20	73.92	56.57
RM3	60.37	52.50	62.00	74.53	56.28
Majority vote	60.75	53.75	63.40	74.22	56.71
Random router	58.39	52.50	61.20	73.46	56.07
UWO (Coste et al., 2023)	61.74	56.25	63.60	74.30	56.43
LASER (Nguyen et al., 2024)	60.50	51.25	62.40	74.00	56.35
w/o offline	60.99	53.75	63.20	74.37	56.64
w/o online	61.61	57.50	64.40	74.68	56.85
BayesianRouter	**63.23**	**58.75**	**66.20**	**75.66**	**57.39**

Table 2: In-distribution and Out-of-distribution performance comparison.

Method	In-distribution	Out-of-distribution				
	Score	Factuality	Precise IF	Math	Safety	Focus	All
RM_0	77.54	75.44	59.22	78.76	85.10	75.61	77.61
RM_1	81.43	84.33	68.44	81.18	96.00	**95.73**	87.65
RM_2	79.51	80.04	67.38	79.03	**97.60**	90.85	85.88
RM_3	81.14	77.64	**70.92**	**90.05**	92.70	92.38	85.34
Majority	83.17	77.74	67.73	85.75	96.50	90.85	85.64
Random	79.80	79.94	65.96	82.80	92.80	87.80	84.48
Ours (w/o CLS)	89.73	84.12	66.67	85.22	96.20	92.99	87.34
Ours (135M)	90.31	84.85	65.60	86.83	96.20	92.07	87.92
Ours (0.5B)	**90.77**	**85.16**	66.31	87.90	95.90	91.46	**88.06**