PAD: Personalized Alignment of LLMs at Decoding-Time

Ruizhe Chen [1,†] Xiaotian Zhang [1,†] Meng Luo [2,†] Wenhao Chai [3,†] Zuozhu Liu [1,]

[1] Zhejiang University [2] National University of Singapore [3] University of Washington

Overview

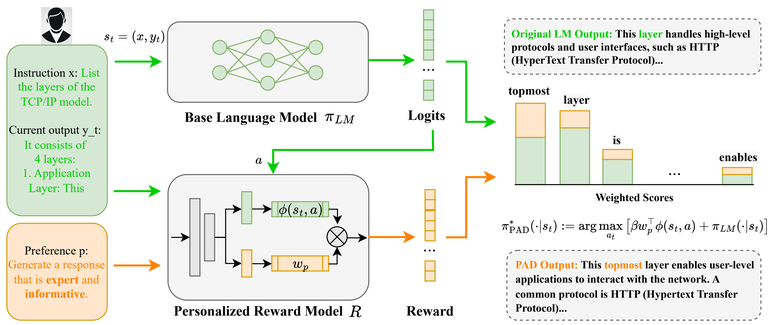
This paper presents Personalized Alignment at Decoding-time (PAD), a novel framework designed to align LLM outputs with diverse personalized preferences during the inference phase, eliminating the need for additional training. By introducing a unique personalized reward modeling strategy, this framework decouples the text generation process from personalized preferences, facilitating the generation of generalizable token-level personalized rewards. The PAD algorithm leverages these rewards to guide the decoding process, dynamically tailoring the base model's predictions to personalized preferences. Extensive experimental results demonstrate that PAD not only outperforms existing training-based alignment methods in terms of aligning with diverse preferences but also shows significant generalizability to preferences unseen during training and scalability across different base models.

Our Main Contributions

The advantages of PAD are as follows: (1) It requires only a single policy model (i.e., the base model) aligned with general preferences (General Policy), eliminating the need for training additional policy models (Training-free). (2) It utilizes only a single reward model (Single Reward). (3) It does not require pre-defined personalized preferences to generalize to preferences not seen during the training phase (Generalizability).

Our contributions can be summarized as follows:

- We propose a novel *personalized reward modeling* strategy that decouples the dynamics of text generation from personalized preferences. This strategy enables the acquisition of generalizable token-level personalized rewards with a single personalized reward model.
- We propose a novel *personalized alignment at decoding-time (PAD)* algorithm that performs guided decoding with the guidance of token-level personalized rewards, while not requiring training additional policy models.
- Extensive experiments demonstrate that PAD outperforms existing training-based methods in aligning with diverse personalized preferences. Furthermore, the results highlight PAD's effectiveness in generalizing to unseen preferences and its model-agnostic scalability.



Method: Personalized Alignment of LLMs at Decoding-Time

Given the personalized preference and the current context, we first calculate the probability distribution of the base model for the next token. Then, we calculate the reward from PersRM combining features of current state and personalized weight. Finally, the next token can be selected based on the weighted scores.

Decoding Phase of PAD

Guided Decoding by Personalized Reward Model

The optimal policy π^*_{PAD} of personalized alignment can be defined as selecting the action for the base model π_{LM} that maximizes the advantage function $Q^*(p, \mathbf{s}_t, \mathbf{a}) - V^*(p, \mathbf{s}_t)$ towards a personalized preference p at each step:

$$\pi^*_{\text{PAD}}(\mathbf{a}|\mathbf{s}_t, p) \propto \pi_{LM}(\mathbf{a}|\mathbf{s}_t) e^{\beta(Q^*(p,\mathbf{s}_t,\mathbf{a}) - V^*(p,\mathbf{s}_t))},$$

where $Q^*(p, \mathbf{s}_t, \mathbf{a}) - V^*(p, \mathbf{s}_t)$ is equivalent to $\mathbf{w}_p^\top \beta \log(\hat{\pi}^*_\theta(\mathbf{a}_t|\mathbf{s}_t)/\hat{\pi}_{\text{ref}}(\mathbf{a}_t|\mathbf{s}_t))$.

Training Phase of the Personalized Reward Model

Definition of Personalized Reward

personalized reward function R can be represented by:

$$R(p, \mathbf{s}, \mathbf{a}) = \mathbf{w}_p^\top \phi(\mathbf{s}, \mathbf{a}),$$

where $\phi(\mathbf{s}, \mathbf{a})$ represents the features of current state and action, and $\mathbf{w}_p$ are weights derived from personalized preference p.

Decouple personalized preferences from the MDP dynamics

The action-value function (i.e., Q function) based on the token-level reward $R^\pi(p, \mathbf{s}, \mathbf{a})$, models the total future reward from $(\mathbf{s}, \mathbf{a})$ under policy π can be expressed as:

$$Q^\pi(p, \mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}[\sum_{i=t}^{T} R(p, \mathbf{s}_i, \mathbf{a}_i)|\mathbf{a}_i \sim \pi(\cdot|\mathbf{s}_i)], = \mathbf{w}_p^\top \mathbb{E}[\sum_{i=t}^{T} \phi(\mathbf{s}_i, \mathbf{a}_i)|\mathbf{a}_i \sim \pi(\cdot|\mathbf{s}_i)] = \mathbf{w}_p^\top \psi^\pi(\mathbf{s}_t, \mathbf{a}_t).$$

Training Objective of the Personalized Reward Model

To obtain the Q^* function, we begin by representing the reward with the policy following DPO,

$$R(p, \mathbf{x}, \mathbf{y}) = \sum_{t=1}^{T} R(p, \mathbf{s}_t, \mathbf{a}_t) = \sum_{t=1}^{T} \beta \log \frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t, p)}{\pi_{\text{ref}}(\mathbf{a}_t|\mathbf{s}_t, p)} + V^*(\mathbf{s}_1).$$

Substitute this relationship with Eq. 1 into the loss function of DPO:

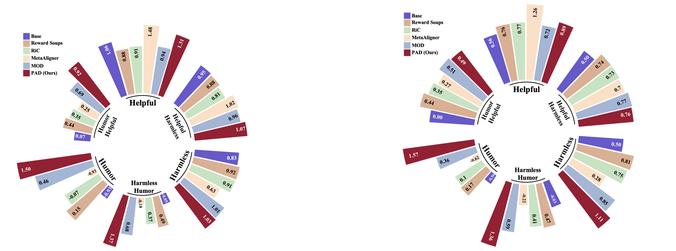
$$\mathcal{L}_{\text{PersRM}}(\pi_\theta, D) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim D}\left[\log \sigma \left(\mathbf{w}_p^\top \left(\sum_{t=1}^{T} \beta \log \frac{\hat{\pi}_\theta(\mathbf{a}_t^w|\mathbf{s}_t^w)}{\hat{\pi}_{\text{ref}}(\mathbf{a}_t^w|\mathbf{s}_t^w)} - \sum_{t=1}^{T} \beta \log \frac{\hat{\pi}_\theta(\mathbf{a}_t^l|\mathbf{s}_t^l)}{\hat{\pi}_{\text{ref}}(\mathbf{a}_t^l|\mathbf{s}_t^l)}\right)\right)\right].$$

Thus we can derive the implicit Q function Q^* with optimized personalized reward model π^*_θ:

$$Q^*(p, \mathbf{s}_t, \mathbf{a}_t) = \mathbf{w}_p^\top \psi^*(\mathbf{s}_t, \mathbf{a}_t) = \mathbf{w}_p^\top \beta \sum_{i=1}^{t} \log \frac{\hat{\pi}^*_\theta(\mathbf{a}_i|\mathbf{s}_i)}{\hat{\pi}_{\text{ref}}(\mathbf{a}_i|\mathbf{s}_i)} + V^*(p, \mathbf{s}_1).$$

Experiment

We conduct our evaluation by focusing on three pre-defined dimensions seen in the training phase: 'harmless', 'helpful', and 'humor'. The results demonstrate that PAD can effectively align with various preferences, outperforming the baselines in terms of achieving a superior frontier. The findings reveal that PAD has achieved substantial improvements for all three objectives. These results demonstrate the superiority of PAD in personalized alignment.



(a) Alignment results on P-Soups dataset. (b) Alignment results on HelpSteer2 dataset.

Figure 1. Alignment results for pre-defined preferences related to harmless, helpful, and humor.

Table 1. Comparison of baseline methods and PAD on predefined preferences. The best result is highlighted in **bold**.

Method	Helpful			Harmless			Humor		Overall	
	Armo	RM	GPT-4	Armo	RM	GPT-4	RM	GPT-4	RM	GPT-4
Base	0.63	1.06	-	**0.97**	0.83	-	-0.93	-	0.32	-
MORLHF	0.31	0.91	14%	0.88	0.84	4%	0.28	82%	0.68	33%
MODPO	0.56	0.89	52%	0.96	0.77	80%	-0.90	72%	0.25	68%
Personalized soups	0.38	-0.72	72%	0.92	0.73	**92%**	-0.30	80%	-0.09	81%
Rewarded soups	0.50	0.87	34%	0.95	0.87	64%	0.14	78%	0.63	59%
RiC	0.54	0.90	40%	**0.97**	0.90	70%	-0.08	76%	0.58	62%
Pref. Promp. (1-dim)	0.56	0.82	70%	0.96	0.87	90%	-0.79	74%	0.30	78%
Pref. Promp. (3-dim)	0.54	0.84	70%	0.93	0.98	87%	-1.28	71%	0.18	76%
MetaAligner (1-dim)	0.47	**1.75**	**79%**	0.90	0.89	71%	-0.74	81%	0.21	77%
MetaAligner (3-dim)	0.55	1.39	66%	0.89	0.54	74%	-0.97	74%	0.32	71%
MOD	0.55	0.93	60%	0.96	0.92	84%	0.38	78%	0.74	74%
Aligner	**0.67**	1.32	72%	**0.97**	0.63	70%	-1.39	12%	0.19	51%
PAD (1-dim)	**0.67**	1.31	74%	0.93	**1.03**	**92%**	**1.50**	**88%**	**1.28**	**84%**
PAD (3-dim)	0.61	0.96	63%	0.98	0.85	87%	0.75	83%	0.85	78%

Contact Us

Paper Code Homepage (Email) WeChat