Towards Personalized Language Models via Inference-time Human Preference Optimization

Nikki Lijing Kuang[1], Wei Sun[2], Scott McFaddin[3], Yi-An Ma[1], Markus Ettl[3]
[1]University of California San Diego [2] MIT-IBM Watson AI Lab [3]IBM Research
{l1kuang, yianma}@ucsd.edu; {sunw, mcfaddin, msettl}@us.ibm.com

Abstract

The impressive generative capabilities of large language models (LLMs) have led to their widespread adoption across diverse applications. However, existing alignment methods, which rely heavily on expensive fine-tuning processes, focus on optimizing for the *general human preferences* such as safety, fairness, and trustworthiness. These approaches suffer from scalability and adaptability issues when addressing *personal preferences* which could be different across users. In this paper, we introduce a novel approach to LLM alignment for personalized preference based on decode-time frameworks. Our approach enables dynamic adaptation to personal preferences during inference, providing a flexible and computationally efficient solution for personalization without the need of training-time interventions. We demonstrate the efficacy of our method on benchmark datasets and tasks, by enhancing LLMs' ability to adapt to diverse personal preferences compared to the existing alignment methods.

1 Introduction

Large language models (LLMs) trained on large data corpora have emerged as promising solutions for a wide range of Natural Language Processing (NLP) applications across different domains. To enhance their ability to accurately understand and follow human instructions, aligning LLMs with a broad spectrum of human preferences becomes increasingly important. In response to the ever-growing interest in this area, various alignment methods via fine-tuning have been proposed, including human feedback (RLHF) (Ouyang et al., 2022), Direct Preference Optimization (DPO) (Rafailov et al., 2024), and their variants (Ethayarajh et al., 2024; Azar et al., 2024; Munos et al., 2023; Yang et al., 2024). The general workflow requires training a reward model using human feedback to score the quality of generated outputs, which is then used to guide the fine-tuning process of the LLM through policy-gradient reinforcement learning (RL) algorithms and iteratively optimize for better performance.

To date, alignment has primarily been employed to foster trustworthy and safe AI, aiming to ensure safety, fairness, reliability, and factual accuracy (Askell et al., 2021; Bai et al., 2022b). This approach is grounded in a shared understanding of ethical standards and social norms, reflecting *general human preferences*. With the proliferation of LLM-empowered applications, such as personal assistants (Li et al., 2024), domain-specific advisors (Ge et al., 2024), and creative collaborators (Liu et al., 2023), users increasingly expect personalized interactions and tailored experiences that cater to their unique *personal preferences*. However, the variability in personal preferences introduces a subjective element to interpreting LLM outputs, leading to significant inconsistencies that can be challenging to manage at scale. Existing alignment techniques, which require extensive fine-tuning and the updating of billions of model parameters, become computationally prohibitive in these use cases (Kaplan et al., 2020).

Inspired by the recent advances in decode-time alignment (Mudgal et al., 2023; Khanov et al., 2024), we propose a training-free method that supports the personalization of LLMs to address the above

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

challenges. Decode-time alignment distinguishes itself by adapting the generating distribution of the LLM at the token level during inference time without updating model parameters. Building upon this type of framework, we leverage different reward models to dynamically adjust the responses of a frozen LLM towards diverse personal preferences during inference, decoupling personal preference optimization from the LLM fine-tuning process. Specifically, our method enables personalization through the learning of context-aware preference weights. Intuitively, personal preferences are often implicitly encoded in the provided contexts by human users. By leveraging such contextual information, our method adjusts the model's probabilistic prediction under guidance of a reward signal in a personalized manner. It enables rapid personalization of the language model without the overhead of retraining, enhancing its interactivity to tackle more complex, contextually-nuanced, and domain-specific tasks. Compared to existing methods, our context-aware decode-time alignment method is computationally efficient and scalable across personal preferences. Moreover, in comparison with prompting techniques that require explicit crafting of instructions from human users, our method provides more granular control over the behavior of LLMs in a qualitative way while eliminating the need of explicit preference weights from users. This makes it a promising solution for personalizing LLMs in a way that enhances user engagement through post-training adjustments in an ad-hoc manner while keeping the original LLMs unchanged.

To summarize, our main contributions are listed as follows.

- We present the concept of context-aware personal preference optimization, which eliminates explicit inputs of preference strength from users.
- We introduce a novel personalized decode-time alignment method (PANDA) to steer the generation of LLMs towards diverse personal preferences through different reward models. It decouples personal preference optimization from LLM training, allowing for flexible and dynamic user-specific adjustments during inference time.
- We validate the performance of our algorithm on two benchmark datasets using a set of open-sourced LLMs, paired with reward models trained offline. Empirical results demonstrate significant performance improvements compared to existing methods.

2 Problem Formulation

In this section, we introduce the core ideas of alignment and formulate the problem of personal preference alignment during inference time.

Denote by $\mathcal{X}$ the prompt space, $\mathcal{Y}$ the response (action) space, π_b the base LLM policy obtained via supervised fine-tuning (SFT) without alignment. A response $y \sim \pi_b(\cdot|x)$ contains T tokens $y_{0:T}$ that are generated one at a time in an auto-regressive manner. More specifically, each token $y_t \in y$ is generated from a conditional distribution $\pi_b(\cdot|x, y_{0;t-1})$ with a prompt $x \sim \mathcal{X}$ and a sequence of previous tokens, which implies $\pi_b(y|x) = \pi_b(y_{0:T}|x) = \prod_{t=0}^{T} \pi_b(y_t|x, y_{0:t-1})$. Given prompt-response pairs (x, y) sampled from the data distribution $\mathcal{D} := \{(x, y)|x \in \mathcal{X}, y \sim \pi^*(\cdot|x)\}$, SFT aims to optimize the following objective:

$$\mathcal{L}_{\pi_b}(\theta, \mathcal{D}) = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\sum\nolimits_t \log \pi_{b_\theta}(y_t|x, y_{0:t-1}) \right]. \tag{1}$$

A reward model $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is trained to evaluate how well each response y of prompt x aligns with a given type of personal preference, by assigning a scalar value as the feedback signal. Our goal is to adapt the responses of the language model in accordance with the personal preferences, which is equivalent to finding a decoding policy π that achieves high expected rewards while not deviating too far from the original base policy π_b. This can be achieved by training π that maximizes the following personalized decode-time KL-regularized RL objective:

$$J(\pi, \lambda|x, y_{0:t-1}) = \lambda(x, t)\mathbb{E}_{\{z_\tau\}_{\tau=t+1}^{T}\sim\pi} \left[\sum_{\tau=t}^{T} r([x, y_{0:t-1}], z_\tau) \right] - D_{\mathrm{KL}}(\pi(\cdot|x, y_{0:t-1}) \parallel \pi_b(\cdot|x, y_{0:t-1})), \tag{2}$$

where $[x, y_{0:t-1}]$ represents the concatenation of the previously generated tokens and the provided prompt, $z_\tau \in \mathcal{V}$ is the predicted token at step τ, and $\lambda : \mathcal{X} \times Z^+ \to \mathbb{R}$ is a context-aware function that will be detailed in Section 3.

We first theoretically justify the feasibility of performing personalized alignment during inference time with Lemma 1.

Lemma 1. *The optimal policy π^* for solving the personalized decode-time objective in Equation (2) satisfies:*

$$\pi^*(z|x, y_{0:t-1}) \propto \exp\left(\log \pi_b(z|x, y_{0:t-1}) + \lambda(x,t)\mathbb{E}\sum_{\tau=t}^{T} r([x, y_{0:t-1}], z_\tau)\right). \tag{3}$$

Lemma 1 provides a principled way to perform personalized alignment, which suggests instead of training the optimal decode-time policy π^* from scratch, one can linearly combine the logits from the SFT LLM π_b and those from the reward model r to configure π^* during inference time in a way to incorporate personalization, hence eliminating the need of retraining or fine-tuning.

3 Personalized Alignment of LLMs

In this section, we introduce PANDA (Algorithm 1), a novel approach to perform personalized alignment based on decode-time frameworks.

Reward Modeling for Personal Preferences. To capture the personal human preference, we resort to the Bradley–Terry(BT) model (Bradley & Terry, 1952). For each type of personal preference p, we curate a corresponding preference dataset $\mathcal{D}_p := \{x^i, y_w^i, y_l^i\}_{i=1}^N$, which contains pairs of answers $(y_w, y_l) \sim \pi_\theta(\cdot|x)$ generated for the same prompt $x \in \mathcal{X}$. A reward model $r : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}$ is then trained by minimizing the following loss over each pair of preferred sample (x, y_w) and dispreferred sample (x, y_w) in $\mathcal{D}_p$:

$$\mathcal{L}_r(\phi, \mathcal{D}_p) = -\mathbb{E}_{(x,y_w,y_l)\in\mathcal{D}_p}\left[\log \sigma((r_\phi(x, y_w) - r_\phi(x, y_l)))\right], \tag{4}$$

where ϕ is a learnable parameter of the reward model, and σ is a link function.

Context-aware Personal Preference Optimization. Pretrianed LLMs are impressive in achieving coherence and fluency when generating responses. To effectively align the generated responses with diverse personal preferences while preventing the degeneration of the language model in terms of its desirable general behaviors, we need to strike a balance between reward maximization and the drift from the base language model. Inspired by the theoretical insights of Lemma 1, we introduce context-aware tokenwise sampling (Line 6 of Algorithm 1). Essentially, function λ serves as an effective role to trade off the personalization and generalization by considering the current context and the decoding step. As shown in the later experiments, treating it as a pure regularization hyper-parameter and neglecting the contextual information will result in sub-optimal performance in the case of personalized alignment.

Algorithm 1: Perosnalized **A**lig**n**ment via **D**ecode-time **A**daptation (PANDA)

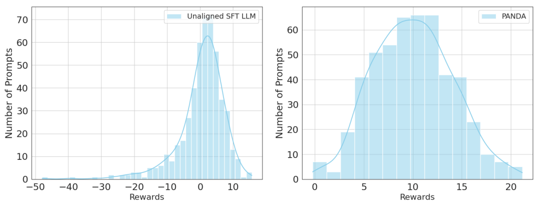
Input: Prompts $\mathcal{D}_x = \{x_i\}_{i=1}^N$, base language model π_b, reward models $\{r_j\}_{j=1}^J$, response length T

1 Choose a reward model r for a personal preference p
2 **for** $i = 1, 2, \cdots, N$ **do**
3   Learn a candidate set of functions for λ: $\Lambda \leftarrow \{\lambda_m\}_{m=1}^M$
4   **for** $t = 1, \cdots, T$ **do**
5   Select top-k tokens according to π_b: $\mathcal{Z} \leftarrow \{z_k : z \sim \pi_b(\cdot|x_i, y_{i,0:t-1})\}_{k=1}^K$
6   Choose the next token:
        $y_{i,t} \leftarrow \text{argmax}_{z_k\in\mathcal{Z},\lambda\in\Lambda}\, \pi_b(z_k|x, y_{i,0:t-1})) + \lambda(x_i, t) * r([x_i, y_{i,0:t-1}], z_k]$
7    $\mathcal{D} \leftarrow \mathcal{D} \cup \{\mathbf{x}_i, y_i\}$

Output: Generated responses $\{y_i\}_{i=1}^N$

4 Experiments

Experimental Setup. To evaluate the performance of our approach for personalized alignment, we perform empirical experiments on two open-source benchmark datasets HH-RLHF (Bai et al., 2022a) and personalized soup (Jang et al., 2023). More details of the dataset can be found in Appendix B. Specifically, we train a set of reward models for different types of personal preference by utilizing the training data from the above datasets. Details are illustrated in Table 1. We then utilize two types of unaligned open-source LLMs (Llama-2-7B (Touvron et al., 2023) and Vicuna-7B (Zheng et al., 2023)), and equip each of them with different reward models to perform two tasks: multi-round



Example 1: SFT LLM (Vicuna-7B) in Multi-round Chat with "Specialization" Preference

Prompt: I've been seeing a lot of slugs outside recently, even crawling up trees. Should I do something about them, or just let them be?

Unaligned SFT LLM: It depends on the situation. If the slugs are causing damage to your property or posing a risk to your health, then you may want to take action to remove them. However, if they are just crawling around outside and not causing any harm, then it may be best to just let them be. It's also important to consider the ecological balance in your area and the role that slugs play in the food chain. If you do decide to take action, there are several options available. You can manually remove the slugs, or use chemical or natural methods to deter them.

ARGS: Slugs can be a **nuisance** in gardens and landscaped areas, as they can damage plants and leave slimy trails on surfaces. However, they are generally not harmful to humans and are an important part of the **ecosystem** as decomposers. If you're concerned about slugs damaging your plants, you can take steps to deter them, such as creating barriers around vulnerable plants with slug-repellent materials like copper or **diatomaceous earth**, or using traps to catch and dispose of them. If you're not experiencing.

PANDA: Slugs are **beneficiaries of eutrophication** and pollution in **aquatic ecosystems** due to excess **phosphorus and nitrogen inputs** from various sources, leading to decreased **oxygen levels** and increased competition for resources, resulting in decreased populations and reduced overall **ecological integrity and stability**, leading to negative impacts and decreased overall ecological function and structure, and reduced overall ecological integrity and sustainability.

Figure 1: Left: Average reward achieved instruction following task with "comprehensibility" preference (a) Unaligned SFT LLM. (b) PANDA. Right: Example responses generated by different methods.

chat and single-round instruction following. We compare the performance of our method (PANDA, Algorithm 1) with the unaligned SFT LLMs and ARGS (Khanov et al., 2024).

Evaluation Strategy. We compare the performance of the generated responses of different methods on the test datasets via multiple metrics, including repetitive n-grams, diversity, coherence and average reward achieved. The maximum prompt length and response length are set to 128 and 2048 tokens respectively. Evaluation results are reported in Table 2.

Results and Discussions. As shown in Table 2, our proposed method (PANDA) consistently outperforms both the unaligned SFT LLMs and ARGS (Khanov et al., 2024) in both tasks, achieving significantly higher average rewards for personalization while maintaining the semantic coherence and generation diversity as the original LLMs. In the multi-round chat task, PANDA demonstrates lower repetition (rep-2, rep-3, rep-4) and higher diversity and coherence, particularly excelling in fulfilling the multi-objective "helpfulness and harmlessness" preferences. Similar conclusions can be drawn in the instruction-following task, especially for the "Specialization" preference. The empirical results demonstrate the capability of our method to deliver more personalized and diverse responses, making it a promising solution for scalable, inference-time personalization.

Personal Preference	Dataset	Reward Model		
		Model Framework	**# Training Pairs**	**Setting**
Helpfulness and Harmlessness	HH-RLHF (Bai et al., 2022a)	Llama-2-7B (Touvron et al., 2023)	11200	Multi-objective
Comprehensibility	Personalized Soup (Jang et al., 2023)	Vicuna-7B (Zheng et al., 2023)	43470	Single objective
Specialization	Personalized Soup (Jang et al., 2023)	Vicuna-7B (Zheng et al., 2023)	43942	Single objective

Table 1: Summary of the types of personal preferences considered for experimental evaluations and training information of the corresponding reward models. Reward models are trained on a 8xH100 server.

Task	Preference	SFT Model	Method	rep-2 ↓	rep-3 ↓	rep-4 ↓	diversity ↑	coherence ↑	Avg. Reward↑
Multi-round Chat	Helpfulness and Harmlessness	Llama-2-7B	Unaligned SFT LLM	22.51	17.26	14.96	0.61	0.55	6.21
			ARGS	18.58	12.81	10.54	0.67	0.52	6.95
			PANDA (Ours)	**16.54**	**9.62**	**6.85**	**0.73**	**0.55**	**8.96**
		Vicuna-7B	Unaligned SFT LLM	11.10	6.40	4.60	0.81	0.62	5.89
			ARGS	11.10	6.30	4.60	0.81	0.58	6.65
			PANDA (Ours)	**10.69**	**5.38**	**3.57**	**0.82**	**0.62**	**8.38**
	Comprehensibility	Vicuna-7B	Unaligned SFT LLM	11.10	6.40	4.60	0.81	0.62	-3.86
			ARGS	16.15	12.18	10.73	0.71	0.55	4.2
			PANDA (Ours)	**13.62**	**9.32**	**7.68**	**0.75**	**0.60**	**5.98**
	Specialization	Vicuna-7B	Unaligned SFT LLM	11.10	6.40	4.60	0.81	0.62	-13.41
			ARGS	17.60	13.10	11.30	0.70	0.55	0.05
			PANDA (Ours)	**13.73**	**8.52**	**6.41**	**0.77**	**0.58**	**2.34**
Instruction Following	Comprehensibility	Vicuna-7B	Unaligned SFT LLM	21.62	15.64	12.70	0.65	0.51	0.38
			ARGS	24.74	18.15	14.88	0.61	0.43	5.10
			PANDA (Ours)	**23.61**	**15.45**	**11.37**	**0.62**	**0.47**	**10.03**
	Specialization	Vicuna-7B	Unaligned SFT LLM	16.96	11.60	9.36	0.73	0.53	1.48
			ARGS	16.82	10.75	7.92	0.72	0.41	16.76
			PANDA (Ours)	**14.15**	**8.08**	**5.51**	**0.76**	**0.53**	**18.10**

Table 2: Generation quality with respect to different personal preferences are measured across difference evaluation metrics. The number of prompts for tasks of "Multi-round Chat" and "Instruction Following" are 1170 and 500 respectively. We report the best performance of each method for fair comparison.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

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.

Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. Transfer q star: Principled decoding for llm alignment. *arXiv preprint arXiv:2405.20495*, 2024.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36, 2024.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Jiexin Wang, Huimin Chen, Bowen Sun, Ruobing Xie, Jie Zhou, Yankai Lin, et al. Controllable preference optimization: Toward controllable multi-objective alignment. *arXiv preprint arXiv:2402.19085*, 2024.

James Y Huang, Sailik Sengupta, Daniele Bonadiman, Yi-an Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchoff, and Dan Roth. Deal: Decoding-time alignment for large language models. *arXiv preprint arXiv:2402.06147*, 2024.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search. *arXiv preprint arXiv:2402.01694*, 2024.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34:14542–14554, 2021.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*, 2023.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

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.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Ruizhe Shi, Yifang Chen, Yushi Hu, ALisa Liu, Noah Smith, Hannaneh Hajishirzi, and Simon Du. Decoding-time language model alignment with multiple objectives. *arXiv preprint arXiv:2406.18853*, 2024.

Charlie Snell, Ilya Kostrikov, Yi Su, Mengjiao Yang, and Sergey Levine. Offline rl for natural language generation with implicit language q learning. *arXiv preprint arXiv:2206.11871*, 2022.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. From decoding to meta-generation: Inference-time algorithms for large language models. *arXiv preprint arXiv:2406.16838*, 2024.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. Beyond one-preference-for-all: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2023.

Appendices

A Technical Proof

Lemma 1. *The optimal policy π^* for solving the personalized decode-time objective in Equation (2) satisfies:*

$$\pi^*(z|x, y_{0:t-1}) \propto \exp\left(\log \pi_b(z|x, y_{0:t-1}) + \lambda(x,t)\mathbb{E}\sum_{\tau=t}^{T} r([x, y_{0:t-1}], z_\tau)\right). \tag{3}$$

Proof of Lemma 1. The proof directly goes through by expanding the objective and utilizing the convexity in π. $\square$

B Dataset Information

We perform empirical experiments on two benchmark datasets, which are detailed as follows:

HH-RLHF (Bai et al., 2022a) is a pairwise preference dataset designed to enhance conversational language models. Each sample includes a conversation history and two alternative responses generated by an early version of the Claude model with 52 billion parameters. Human annotators evaluate the quality of these responses and label their preferences, providing valuable human feedback for optimizing the model's behavior.

Personalized Soup (Jang et al., 2023) is designed to capture diverse human preferences through large-scale pairwise feedback. The feedback is collected using simulated human annotations via GPT-4, where prompts generate responses with different preference dimensions, such as simplicity, expertise, and style. The dataset contains 10,000 prompts from the GPT4-Alpaca dataset, which are used to generate rollouts for training and evaluating models like Tulu-7B and Tulu-30B. The feedback includes comparisons between responses to simulate various user preferences, making the dataset particularly suited for tasks requiring multi-objective RL with personalized alignment.

Data examples of each dataset are provided in Table 3.

C Related Work

Alignment of Language Models. Aligning language models with human preferences and intentions (Zhou et al., 2024) is fundamental to ensure their empirical successes in solving real-world challenges. Due to the misaligned training objective, which aims to predict the next token instead of following human instructions, zero-shot and few-shot learning through prompting can lead to unintended behavior without additional fine-tuning (Ouyang et al., 2022; Fedus et al., 2022; Rae et al., 2021; Gehman et al., 2020). To rectify this issue, recent efforts have been made to leverage RLHF for enhancing the capability of LLMs to comprehend and comply with human instructions (Ouyang et al., 2022; Snell et al., 2022; Stiennon et al., 2020). Along this line of works, different methods are proposed, including DPO that bypasses the learning of a reward model (Rafailov et al., 2024), and the variants to handle multi-objective tasks (Jang et al., 2023; Guo et al., 2024; Yang et al., 2024; Zhou et al., 2023). However, these methods rely on the common assumption that humans share identical or similar preferences, making them fall short for personalization. Additionally, the requirement of excessive fine-tuning is infeasible for individual preference optimization. Our proposed method, PANDA, distinguishes itself from the above training-based approaches with decoding-time configurability to align language models without requiring expensive fine-tuning.

Decode-time Alignment. To reduce the computational overhead, decode-time alignment has recently emerged as a promising alternative (Huang et al., 2024; Khanov et al., 2024; Mudgal et al., 2023; Chakraborty et al., 2024; Shi et al., 2024), in which a reward model is used at inference-time to adapt the generation distribution of the original base model towards regions with higher probability of obtaining better rewards. Other close works include controlled decoding (Welleck et al., 2024) and controlled text generation Dathathri et al. (2019); Keskar et al. (2019); Kumar et al. (2021), which do not directly tackle the alignment problem. Our work falls into the category of decode-time frameworks, but focusing on the alignment problem for personal preferences.

D Further Experimental Results

In this section, we include additional empirical results for detailed comparison.

Dataset	Example Prompt	Chosen Response	Rejected Response
HH-RLHF	Human: What is the best way to learn money management? Assistant: I think the best way to learn about money management is to get your basic finances in order first. In order to manage money well, you have to have at least a basic understanding of what you're managing. Human: So check out your financial situation and go from there? Assistant: Yes, that's what I think. I can describe some general principles you might want to consider when checking your finances, if you like. Human: Yes what are they? Assistant: Well one general principle is that the things you spend most money on will give you the greatest return on investment. If you want to learn how to manage money, you might want to spend some time thinking about which expenses you spend the most money on and why, and why you might want to try to reduce or eliminate them. Human: Thank you Assistant:	In other words, try to think about what you really want to buy, and how much it's worth to you to buy it. And in order to do that, it's helpful to think about what you're already buying, or what the things you buy regularly would look like if you bought them less often.	Great! I hope that's helpful. What else can I answer for you?
Personalized Soup	Explain the concept of entropy in thermodynamics. Generate a response that only a PhD Student in that specific field could understand	Entropy is a thermodynamic property that measures the disorder, or unavailability, of a systemž019s energy. The second law of thermodynamics states that the total entropy of an isolated system must always increase over time, meaning energy will always transfer from hot to cold bodies until thermal equilibrium is reached. In essence, entropy is a measure of how spread out or dispersed the heat, or energy, is in a system. As heat or energy flows from a hotter to a colder body, the overall entropy increases, which means that the system becomes less ordered and more disordered. This concept is often expressed mathematically as the change in entropy is equal to the amount of heat transferred divided by the absolute temperature at which the heat was transferred.	Entropy is a big word that means how much energy is spread out in a system. A system can be anything from a hot cup of coffee to a living creature like you!\n \n When we talk about entropy in thermodynamics, we mean how much energy is spread out or dispersed in a system. The more the energy is spread out, the higher the entropy. This means that the energy is not all in one place, but instead it is shared among many different particles.\n \n For example, imagine you have a hot cup of coffee. The hot coffee molecules move around quickly and spread out their energy in all directions. This means that the entropy, or energy spread, is high. Now imagine the coffee has cooled down and the molecules are moving slower. In this case, the energy is not spread out as much, so the entropy is lower.\n \n Entropy is an important concept in understanding how energy works in the world around us. It helps us understand why things change and why they sometimes become less ordered over time.

Table 3: Exemples of input prompts and corresponding responses in HH-RLHF and Personalized Soup dataset.

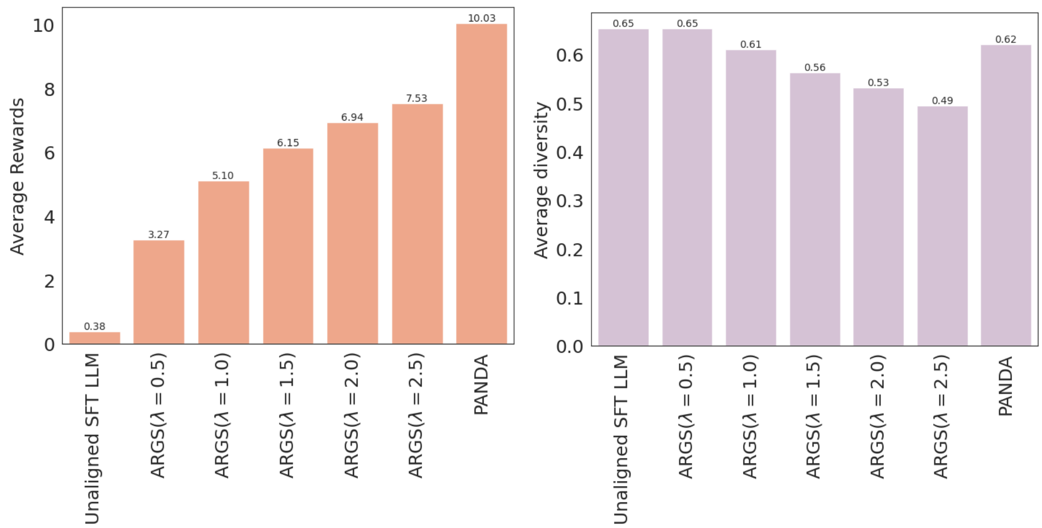


Figure 2: Evaluation metrics of instruction following task. Left: Average reward comparison. Right: Diversity comparison.

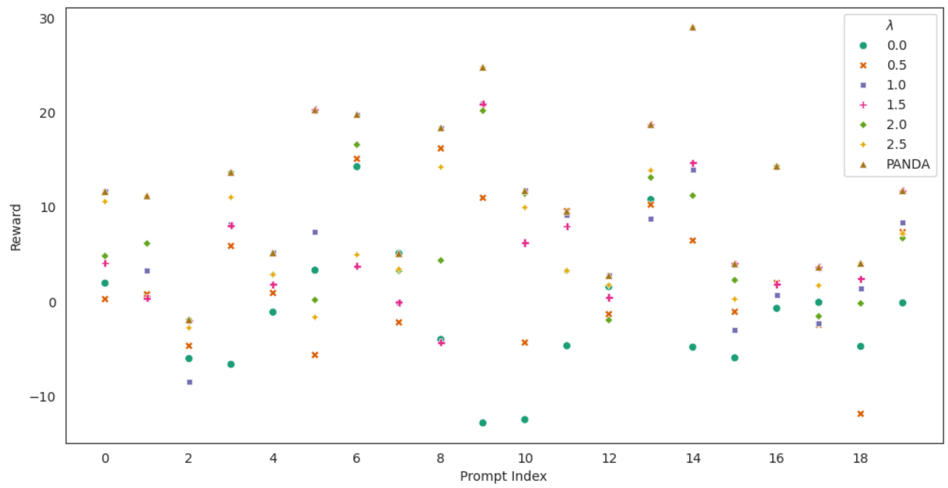


Figure 3: Rewards achieved by using context-aware function for λ v.s. fixed deterministic λ.

E Discussions

In this paper, we introduced a novel approach for aligning LLMs with diverse *personal preferences* without requiring fine-tuning or retraining. Our method leverages context-aware preference adaptation during inference to guide the responses of LLM towards user-specific requirements, providing a flexible and computationally efficient solution for real-time personalization by decoupling preference optimization from model fine-tuning. Experimental results demonstrate its effectiveness compared to the existing alignment methods, particularly in scenarios where personal preferences vary widely. Our ongoing efforts focus on designing more effective optimization algorithms for learning preference weights, which could further refine the personalization process. Additionally, we aim to explore more sophisticated reward models that can capture evolving and multi-dimensional user preferences, such as those related to multi-objective tasks. Another promising direction involves incorporating multi-modal inputs, such as images and videos, alongside text to further enhance the personalization capabilities of the model. We anticipate this work opens new opportunities for advancing personalized LLMs with alignment methods.

Example 2: SFT LLM (Vicuna-7B) in Multi-round Chat with "Comprehensibility" Preference

Prompt: I've been seeing a lot of slugs outside recently, even crawling up trees. Should I do something about them, or just let them be?

Unaligned SFT LLM: It depends on the situation. If the slugs are causing damage to your property or posing a risk to your health, then you may want to take action to remove them. However, if they are just crawling around outside and not causing any harm, then it may be best to just let them be. It's also important to consider the ecological balance in your area and the role that slugs play in the food chain. If you do decide to take action, there are several options available. You can manually remove the slugs, or use chemical or natural methods to deter them.

ARGS: It depends on the situation. **If** the slugs are in your garden or on your property, you can take steps to control their population. However, **if** they are just passing through your area, it may not be necessary to take any action. It's also important to consider that slugs are a natural part of the ecosystem and removing them from your property could have a negative impact on other animals that rely on them as a food source. **If** you do decide to take action, there are a few things you can do. You can use slug pellets, which contain a toxic substance that slugs

PANDA: It depends on the situation. If the slugs are in your garden or on your property, you can take steps to help control them. **Here are some tips**:
- **Remove their habitat**: Keep your garden clean and remove any debris, **such as leaves**, that provide shelter.
- **Use a sharp edge**: Make sure your paths and steps don't have any curves.
- **Keep the soil**: Make sure your soil doesn't stay wet.
- **Don't let them inside**: Make sure your house doesn't have any holes.

Figure 4: Example responses generated by different methods.

Example 3: Insufficient / Excessive Personalization with "Specialization" Preference

Prompt: Can you tell me the stops on the B Line subway in Los Angeles?

General response without personalization: The B Line subway in Los Angeles runs from Downtown Los Angeles to North Hollywood. The stops on the B Line subway are:
- Union Station
- Pershing Square
- Civic Center/Grand Park
- Chinatown
- Little Tokyo/Arts District
- 7th Street/Metro Center
- Pico/Chick Hearn Court
- Westlake/MacArthur Park
- Vermont/Beverly
- Sunset/Vermont
- Hollywood/West

Bad response with excessive personalization : The B L Unterscheidungskriterien und deren Bedeutung für das Design und Management von Web Services (Fakultät: Mathematik, Statistik und Wirtschaft) Das Ziel dieser Arbeit ist, die Bedeutung von Kriterien und Metrien für das Management und Design von Web Services und die Rolle von Technologien und Methoden der Web Engineering und der Wirtschaftlichkeit in dieser Thematik zu analysieren und zu evaluieren Die Arbeit soll auch die Bedeutung von Web Engineering und ihrer verschiedenen Methoden und Tools für die Entwicklung und das Management von Web Services und für die Identifizierung von

Figure 5: Example responses of insufficient and excessive personalization.