Replication of "*Personalizing Reinforcement Learning from Human Feedback with Variational Preference Learning*"

Peter Phan
College of Information and Computer Science
University of Massachusetts
Amherst, MA 01003
pkphan@umass.edu

Abstract

Reinforcement Learning from Human Feedback (RLHF) methods have grown to be a cornerstone tool for aligning models to human-centric values. Current RLHF techniques, which rely on unimodal reward models such as the Bradley-Terry-Luce (BTL) model, struggle to account for the complex, multi-modal nature of human preferences in diverse populations. This work explores a proposed solution to these issues coined as Variational Preference Learning (VPL). VPL emphasizes its ability to infer and adapt to individual user contexts through a latent variable formulation. We investigate the core hypothesis that VPL can effectively capture multi-modal reward functions, demonstrating its capacity to recover diverse preference distributions more accurately than existing methods. Experimental results show that VPL significantly outperforms the BTL baseline in recovering complex, multi-modal reward functions. These findings suggest that VPL is a promising approach for pluralistic alignment, offering improved alignment to diverse values and preferences in RLHF systems.

1 Introduction

As artificial intelligence (AI) models become increasingly commonplace and influential to our decision-making, it is crucial that these models are aligned to its users' values – both for safety and application efficacy. Reinforcement Learning from Human Feedback (RLHF) has shown to be an effective method in the domains of natural language processing (NLP) and robotics for aligning models to human preferences [Ouyang et al., 2022, Leike et al., 2018]. Current RLHF approaches presumes that all users in a given population share the same values [Ouyang et al., 2022]. This is a problematic assumption to make for diverse populations where moral, social, and political preferences can vary significantly. Furthermore, recent work has shown that current RLHF methods may ignore preferences of minority groups, making them sub-optimal and unfair for certain or all groups. This work will focus on a system for *pluralistic alignment* [Sorensen et al., 2024] of models to human preferences, which aims to democratize RLHF to align with the wide range of human values in a diverse population. Given the recent novelty of this problem, it is important to verify the robustness and effectiveness of newly proposed solutions.

Current RLHF approaches use the Bradley-Terry-Luce (BTL) [Bradley and Terry, 1952] model for learning reward models that infer human preferences. The BTL model typically makes the 'unimodal' assumption that all human preferences are derived from a single reward model. In addition to diverse values, human preferences may also be irrational and influenced by hidden context [Siththaranjan et al., 2024]. Therefore the BTL assumption fails in scenarios where preferences are multi-modal due to different fundamental utility functions. Take the case in Figure 1 for instance where one group of users may prefer short and concise responses while another group of the same size prefers long and detailed responses. Maximum likelihood estimation under the unimodal BTL model will

learn a reward function that averages both of these preferences. Training a policy model, such as a large language model (LLM), under this reward function will then result in an optimized policy for producing medium-length responses with moderate details—a behavior that is sub-optimal for both sub-populations of preferences. Therefore, vanilla RLHF methods fail at the task of pluralistic alignment to diverse values—motivating the need for a system designed for multi-modal reward modeling such as Variational Preference Learning (VPL) [Poddar et al., 2024].

The notion of hidden user context Siththaranjan et al. [2024] can be used to explain the variations we see in human preferences. Therefore, it is beneficial to be able to infer and adapt to individual user's context in order to accurately model each user's utility. With this motivation, VPL builds on techniques from variational inference and phrases RLHF as a latent variable problem. The intuition behind VPL is to use a learned variational encoder on a set of preference annotations from some rational user to infer the latent distribution over hidden user context, and a latent conditional reward model to recover the true multi-modal preference distribution. The authors of VPL empircally showed that their system effectively learns an accurate distribution of reward functions from corpora of different preferences produced by diverse users in various controlled and LLM environments.

In this work, we will narrow in on one of the core contributions of VPL, which is about learning latent-conditioned reward models for modeling multi-modal preferences. We will validate VPL's hypotheses for answering the research question: How well does VPL capture multi-modal reward functions of varying complexities? These hypotheses are summarized below.

- VPL is able to recover a distribution of reward functions that closely resembles the ground truth reward distribution. Particularly, we expect to see multiple preference modes corresponding to the modes of the ground truth distribution of preferences.
- VPL can scale and produce a reward model that can model a large number of different preferences. Meaning, a learned VPL reward model on a population with many different preferences can be conditioned to produce the reward function for every unique preference.

2 Related Work

2.1 Reinforcement Learning from Human Feedback (RLHF)

The problem setting of the VPl paper is about RLHF from binary human preferences using the BTL model [Bradley and Terry, 1952]. This sub-field of reinforcement learning (RL) and robotics is known as Preference-based RL (PbRL) [Hejna and Sadigh, 2023]. The VPL paper focused on the specific RLHF framework described by [Christiano et al., 2023]. The applications of the RLHF framework includes training robots and finetuning large language models for alignment [Christiano et al., 2023, Ouyang et al., 2022]. As such, VPL is applicable to any binary preference-based learning method [Poddar et al., 2024].

2.2 Personalized RLHF

Prior works in using non-BTL models aims to account for human irrationality and preference uncertainty. The work done in VPL is less focused on human irrationality and more about the divergence of rational preferences between different humans [Laidlaw and Russell, 2021]. That is not to say VPL can not also be used to consider irrationality. By its design, VPL attempts to model and align to all behaviors in its training which may include irrationalities.

Other works approach the same problem as VPL through more societal lens [Kirk et al., 2023]. They reiterate the need for personalization and propose datasets with diverse and explicit annotations. Unlike the VPL paper, these works do not present a technical solution for modeling diverse preferences in a population.

Some other works address the issues of alignment to conflicting objectives through methods such as multi-objective RL or Pareto-optimal optimization [Boldi et al., 2024]. [Jang et al., 2023] tries learning independent reward models for different users and [Dai et al., 2023] categorizes different preferences as "helpfulness" and "harmfulness". In contrast, VPL does not aim to reconcile diversity in preferences but rather solve the model misspecification issue by learning reward models that can infer hidden contexts and specialize to a particular user.

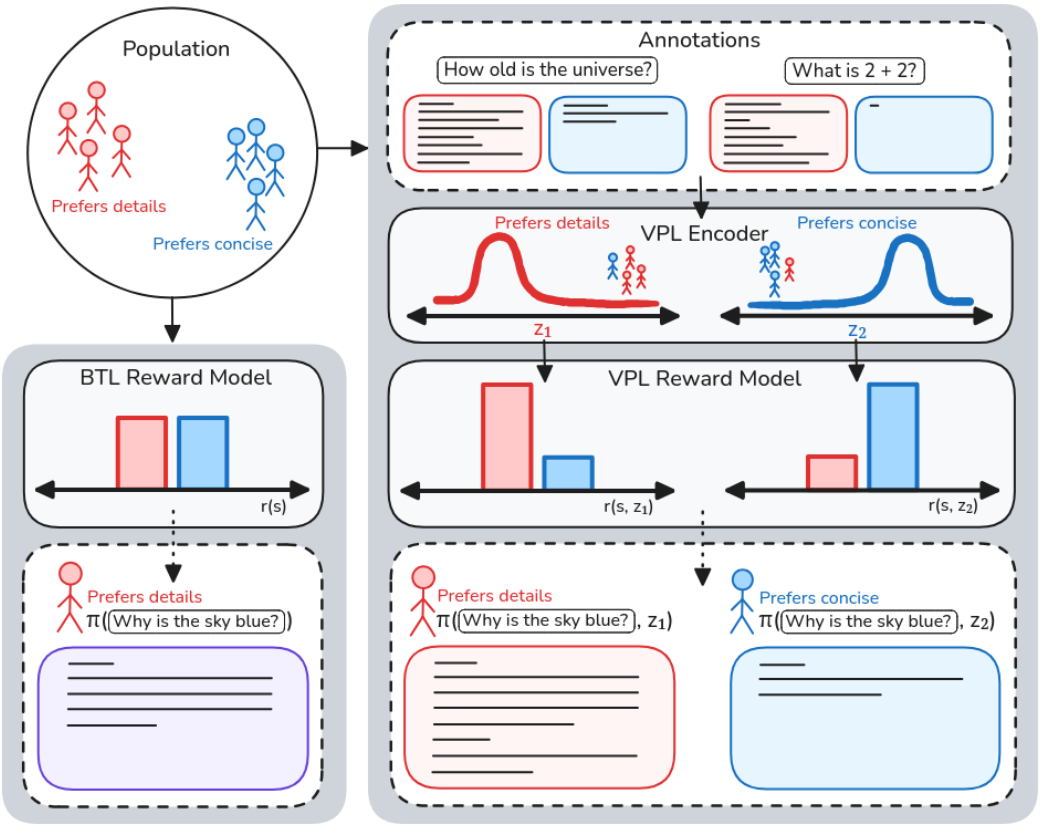


Figure 1: In the depiction, a population of users have equally opposing preferences over the level of detail provided by a large language model response. The RLHF approach on the left, using a unimodel BTL reward model, gives equal rewards to both detailed and precise responses. In contrast, the VPL approach to personalized RLHF on the right depicts a reward model being aligned with its corresponding users' preference. VPL first queries the user to annotate their preferred responses for a few examples. Based on the distribution of their annotations, VPL encodes the user's preference and conditions a reward model to accurately predict rewards accordingly. The downstream policy is also conditioned on the latent preference, enabling it to be steered to the given user's preference.

There are some works that model diversity in preferences by using extensive user information to map a user to a specific category in a fixed set of preference representations [Li et al., 2024]. In comparison, VPL does not require access to presonal user information and does not make the inplicit assumption that all members of a demographic shares the same preferences. Instead, VPL bases its preference assumption of the user on a few preference annotations from the user at test-time.

Works that do personalization at test-time do so outside the context of reward learning. This includes [Zhao et al., 2024] which uses in-context learning to learn preferences for discrete multiple choice answer to a prompt. VPL learns a reward model which has the benefits of explainability and transferability to active learning and latent-conditiond policy learning.

The most similar work to VPL is Distributional Preference Learning (DPL) [Siththaranjan et al., 2024]. DPL focuses on a similar problem statement for acounting for hidden context in RLHF. DPL's approach is to model the mean and variance of the reward distribution across users. By doing so, DPL is able to model the uncertainty of the inferred model distribution but cannot model the reward function of a specific user. VPL differs by learning a conditional reward model that exhibits different user-specific reward models according to the user preference latent variable z that it is conditioned on. This allows VPL to have more explainable rewards and model specialization for each individual user. In their paper, the authors of VPL empircally showed that VPL is significantly more accurate at prediction rewards than DPL.

3 Methodology

3.1 RLHF with BTL

VPL builds on the reward learning framework based on the BTL choice model. RLHF using the BTL model has two core phases. The first step infers a reward function from human-provided labels of binary preferences. The second step is to use RL to train a policy that takes actions that maximizes the rewards inferred in the prior step. For the purpose of this replication paper, we will present the framework abstractly. further describes their approach for specializing this framework for both LLM and robotics environments.

A Markov decision process (MDP) is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \eta, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P}$ is the transition dynamic, $\eta(\cdot)$ is the initial state distribution, and $\gamma \in (0, 1)$ is the discount factor. Note that we do not have access to the underlying reward function. Instead, we have access to a human user $h \in H$ that takes takes in pairs of states, (s_A, s_B), and returns a binary preference $y = \mathbb{I}(s_A \succ s_b)$ according to their internal reward function $r(s)$. Given a population, H, and a dataset of annotated prefences from that population $\mathcal{D} = \{(s_A^i, s_B^i), y = \mathbb{I}(s_A^i \succ s_b^i)\}_{i=1}^N$, the traditional RLHF approach learns a reward function $r_\phi(s)$ parameterized by ϕ, using a maximum likelihood objective (MLE) on the preferences. Here, the likelihood of the preference for a pair of states, $p_\phi(y|s_A, s_b)$ is defiend by the following BTL model:

$$p_\phi(y = 1|s_A, s_B) = p_\phi(s_A \succ s_b) = \frac{e^{r_\phi(s_A)}}{e^{r_\phi(s_A)} + e^{r_\phi(s_B)}} \tag{1}$$

Note that the states s_A, s_B can be generalized to trajectories in the robotics setting or text responses in the LLM setting. After learning the reward model, r_ϕ, the recovered reward function is used in the second phase of RLHF to train or finetune a policy $\pi_\theta(a|s)$. This policy, $\pi_\theta = \arg\max_\theta \mathbb{E}_{\pi_\theta}[\sum_t \gamma^t r_\phi(s_t)]$, is trained to maximize the acquired rewards in its given environment using standard RL algorithms such as PPO, Soft Actor-Critic, or implicit q-learning. It has been shown that this probabilistic formulation of the BTl model is successful at accounting for noise in preferences. However, it does not account for hidden context and conflicting human preferences, thus not allowing the underlying reward models and policioes to be personalized to individual users.

3.2 RLHF with VPL

Notice that the standard BTL formulation is based on the assumption that all annotators in a population H share a single underlying reward function modeled by $r_\phi(s)$. To modal diverse preferences, VPL frames the multi-modal reward learning objective as a latent variable problem. The latent variable z represents the hidden context that affects the user's, h, underlying reward function. This latent variable corrects the reward function to account for the user's hidden context at test time, thus being more indicative of the user's current preference. VPL defines the latent-conditional reward, $r_\phi(s, z)$, to be a function of the given state s and given latent variable z. Similar to the BTL approach, VPL defines the likelihood of the preference for a pair of states, $p_\phi(y|s_A, s_b)$ as the following:

$$p_\phi(y = 1|s_A, s_B, z) = p_\phi(s_A \succ s_b|z) = \frac{e^{r_\phi(s_A, z)}}{e^{r_\phi(s_A, z)} + e^{r_\phi(s_B, z)}} \tag{2}$$

Given a dataset of preference labels, the maximum likelihood objective for this model, $\max_\phi \mathbb{E}_{s_a, s_B, y\sim\mathcal{D}}[\log p_\phi(y|s_A, s_B)] = \mathbb{E}_{s_A, s_B, y\sim\mathcal{D}}[\log \int p_\phi(y|s_A, s_B, z)p(z)dz]$. Notice that the integral over the latent variable z is intractable. To solve this, VPL formulates an evidence lower bound (ELBO), $\mathcal{L}(\phi, \psi)$, for the intractable marginal $\log p_\phi(y|s_A, s_B)$. VPL introduces a variational posterior approximiation $q_\psi(z|\{(s_A^i, s_B^i, y^i)\}_{i=1}^N)$ conditioned on a set of annotations from a given user h. This posterior, q_ψ, refers to the VPL encoder shown in Figure 1. Altogether, the ELBO objective is defined as follows:

$$\mathop{\mathbb{E}}_{\substack{h\sim H \\ \{s_A^i, s_B^i, y^i = h(s_A^i, s_B^i)\}_{i=1}^N \sim \mathcal{D} \\ \{s_A, s_B, y = h(s_A, s_B)\} \sim \mathcal{D}}} \left[\mathop{\mathbb{E}}_{z\sim q_\psi(z|\{(s_A^i, s_B^i, y^i)\}_{i=1}^N)} \log p_\phi(y|s_A, s_B, z) - D_{\text{KL}}(q_\psi(z|\{(s_A^i, s_B^i, y^i)\}_{i=1}^N) \| p(z)) \right] \tag{3}$$

In summary, the objective first samples a user $h \sim H$ and a set of annoations from this user $\{(s_A^i, s_B^i, y^i = h(s_A^i, s_B^i))\}_{i=1}^N$. These annotations are used to infer the latent variable z from the

posterior $q_\psi(z|\{(s_A^i, s_B^i, y^i)\}_{i=1}^N)$. So together, the objective optimizes a maximum preference likelihood objective, $\log p_\phi(y|s_A, s_B)$, and a regularization term $D_{\text{KL}}(q_\psi(z|\{(s_A^i, s_B^i, y^i)\}_{i=1}^N)\|p(z))$ against the prior $p(z)$. From this objective, q_ψ encodes a set of user-provided annotation to a latent distribution z, then the reward function, $r(s, z)$, learns to explain this annotated preference data from an encoded preference. Algorithm 1 gives an overview of the implementation of the VPL system.

Algorithm 1 Learning Multimodal Reward Functions using VPL [Poddar et al., 2024]

Require Preference Data $\mathcal{D} = \{(s_A^i, s_B^i, y^i)\}_{i=1}^N$
Require Encoder E, Reward model R, prior $p(z)$
while not done **do**
  Sample batch $B \sim \mathcal{D}$
  Compute $\mu_B, \sigma_B = E(B)$ ▷ Latent distribution
  Sample $z \sim \mathcal{N}(\mu_B, \sigma_B)$ ▷ Latent variable
  Append z to B: $\{(s_A, s_B, y)\} \rightarrow \{((s_A|z, s_B|z), y)\}$
  Compute rewards: $r_{s_A} = R(s_A|z)$, $r_{s_B} = R(s_B|z)$
  Comupte reconstruction loss: $\mathcal{L}_{\text{recon}} = \text{cross entropy}(\sigma(r_{s_A} - r_{s_B}), y)$ ▷ Equation 2
  Compute KL-loss: $\mathcal{L}_{\text{KL}} = \beta \cdot D_{\text{KL}}(\mathcal{N}(\mu_B, \sigma_B)\|p(z))$
  Compute total loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}}$ ▷ Equation 3
  Update E and R by optimizing $\mathcal{L}_{\text{total}}$
end while

4 Research Design

As described in section 3, one of the core step in VPL is to leverage a learned latent-conditioned reward model to condition policies to adapt to different preferences. Recall that the research question we then want to answer is about the effectiveness of this approach for representing multi-modal rewards. The first hypothesis will focus on a comparison between the BTL and VPL approaches where we will show the difference in the learned reward models of the two paradigms. The second hypothesis will accees the VPL system's capability to scale to higher complexities of multi-modal reward functions.

4.1 Hypothesis 1

Once we have our latent-conditioned reward model, we want to know if it will actually produce the rewards of multi-modal reward functions. The first hypothesis is that conditioning the reward on a latent variable of the inferred preference will lead to rewards that are more aligned with that preference than the traditional BTL model.

To test this hypothesis we will create a didactic experiment. Let us consider a mixture of M annotators providing different preferences, where each annotator i has a reward function specified by a gaussian $\mathcal{N}(\mu_i, \sigma_i)_{i=1}^M$ that they use to assign binary preferences. We will sample preferences from this mixture of Gaussians as follows:

$$p(s_A \succ s_B|i) = \frac{e^{r_i(s_A)}}{e^{r_i(s_A)} + e^{r_i(s_B)}}; \text{where } e^{r_i} \sim \frac{1}{\sigma_i\sqrt{2\pi}} e^{\frac{1}{2}(\frac{x-\mu_i}{\sigma_i})^2} \tag{4}$$

Using this formulation, we can simulate multi-annotator preferences by sampling an annotator from this mixture distribution and then assigning binary preferences according to the sampled reward function. We will then train VPL to recover the underlying distribution over reward functions. We will also compute the underlying distribution over reward functions using the BTL model. Once we have these results we can compare the recovered reward distributions with the original distribution denoted by the M annotators. Our hypothesis states that we should expect VPL to recover a distribution that is closer to the ground truth than BTL is able to.

4.2 Hypothesis 2

We now want to explore how VPL behaves with varying complexity of preferences. One way to increase this complexity is to increase the number of possible different preferences we want to align to. The second hypothesis is that the high dimensionality of the latent preference variable will allow VPL to scale and align to many differerent preferences.

To test this hypothesis about VPL's effectiveness at scaling to many modes of preferences, we will create an experiment with 10 and 120 underlying possible preferences. The first experiment with 10 preferences will take place in the Maze-Navigation environment. There will be 10 underlying locations in the maze that the user could prefer the policy to navigate to. The challenge for VPL will be to disambiguate the given user's preference among the 10 possible cases and correctly condition the policy to navigate to the desired location. We will perform the same experiment with the BTL model. We will then compare the success rates or reaching the desired locations with respect to each approach.

We will further test this hypothesis with another experiment where we will scale the number of preferences to 120. This experiment will take place in the Habitat-Rearrange environment. In this environment, the objective is to control the Mobile Manipulator robot to pick up a bowl and move it to the user's preferred location. The five possible locations are 'desk', 'room', 'dining', 'coffee table', and 'sofa'. The user will have a ranking preference over the 5 locations for where they would prefer the bowl. In total this will give us $5! = 120$ possible different preferences. For the purpose of this experiment, the problem is reduced to a discrete one-step problem, where the robot only has to reason about the best possible location to put the bowl—instead of completing a full trajectory to the desired location. At test time, we will query the agent for the location that maximizes the inferred reward from the its reward model–denoting the location it wishes to move the bowl to. Similar to the Maze-Navigation experiment we will record and compare the success rate of VPL and BTL at moving the bowl to the desired location.

5 Experimental Results

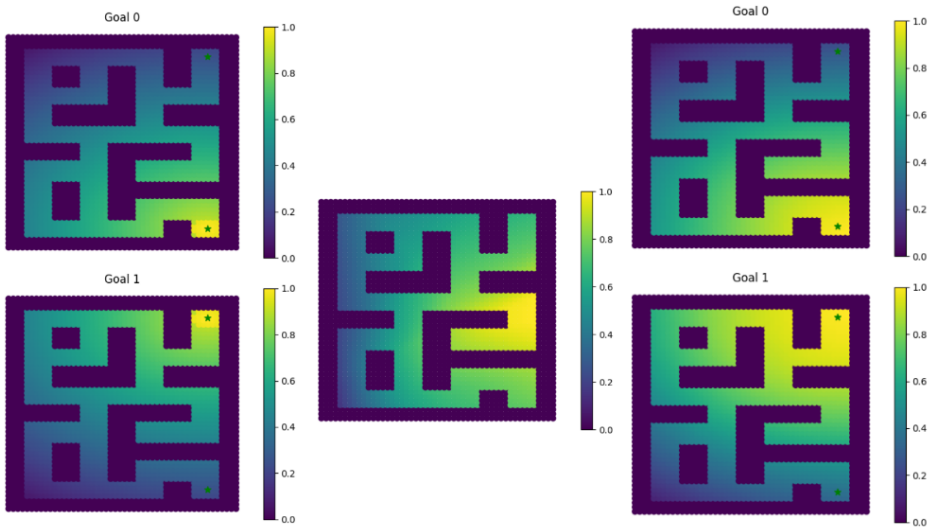
First we depict a comparison of the behavior of BTL and VPL on the Maze-Navigation environment. We see that training a reward model using BTL with the two preferences shown in **??**a will produce a reward model that averages the underying two rewards as shown in **??**b. We can see that this reward model is inaccurate to either of the ground truth rewards. Using the latent encoded preferences of Goal 0 and Goal 1, VPL is able to recover both ground truth reward models as shown in Figure 2.

We trained VPL as described in Section 4.1 to recover the underlying distribution over reward functions. As expected in theory, Figure 2 empirically shows that that standard RLHF with BTL averages over the different preference modes since it can only represent a single reward function. We also compared against prior work that accounts for hidden context in RLHF, particularly DPL [Siththaranjan et al., 2024] which can learn the uncertainty in reward functions due to hidden context. Figure 2 shows that DPL is unable to accurately disambiguate different preference modes. In comparison, VPL is able to infer the underlying context using the approximate latent posterior q_ψ and recover the individual reward modes through the latent-conditional reward function $r(s, z)$.

Figure 4a shows that VPL is able to navigate to the individual goals with a higher success rate. Whereas the baseline BTL model achieved a success rate of 0.1 since it collapses to one of the ten user modes. In Figure 4b we observe that VPL significantly outperforms the baselines in inferring the 120 user preferences, and steering the robot policy accordingly.

6 Conclusions

From didactic and simulated experiments we generated preferences using multi-modal reward functions and evaluated VPL's ability to recover these underlying rewards. As shown in Figure 2, 3, and 4, we see that the standard BTL baseline trend towards averaging the underlying rewards which results in a far from optimal reward model. These experiments strongly indicates that VPL is capable of inferring the hidden user context using the latent variable formulation and accurately recovers the multi-modal reward distribution.



(a) Ground truth (b) BTL (c) VPL

Figure 2: Here, the gradient indicates the reward value given at each location of the maze. Ground truth preferences (a) show that annotators prefer the robot navigate to two different goals. Unimodal BTL (b) averages over the two preference modes. VPL (c) accurately reconstructs diverse preferences, and learns z conditioned reward functions for each user.

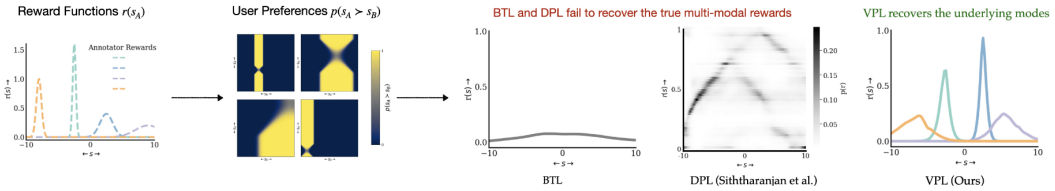
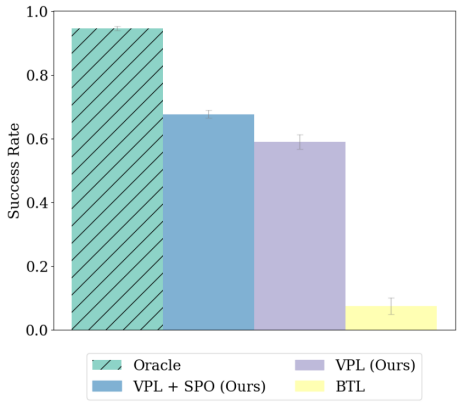


Figure 3: Didactic experiment comparing standard BTL [Bradley and Terry, 1952], DPL [Siththaranjan et al., 2024], and VPL [Poddar et al., 2024]. Four Gaussian reward functions generate different binary preference data and are depicted in the first 2 plots. The third plot shows traditional BTL averages the different modes, and the fourth plot shows that DPL captures the uncertainty in the rewards due to the multi-modality but cannot accurately predict the true modes. In the fifth plot we see that VPL can infer the hidden latent and recover individual distributions that resemble that of the ground truth reward functions.

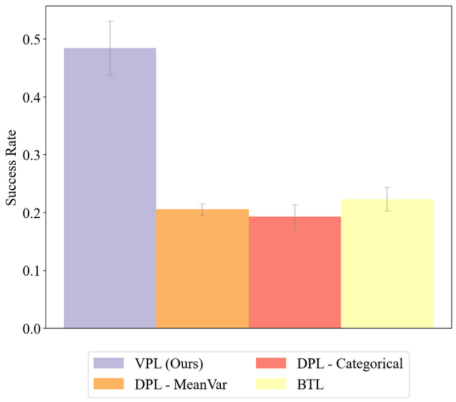
The evidence for this conclusion is a mostly internal validity. The results of the BTL baselines is supported in theory and by the simulation of our experiment. This makes it evidence in the realm of internal validity. The evidence for VPL's multi-modal reward function representation is a wide range of simulations. The experiments show that VPL is successful at recovering multi-modal rewards– especially compared to prior works–in the selected environments we performed the experiments in.

6.1 Future Work

Additional experiments comparing VPL with other systems in more settings would improve the statistical correlation of VPL and multi-modal reward modeling.



(a) VPL scales to Maze-Navigation task with ten modes of user preferences. BTL expectedly averages the modes and fails to learn. We also see the benefits of scaling rewards across this domain well, where VPL + SPO performs better than VPL.



(b) We compare the performance of baselines and VPL on a Habitat-Rearrange environment with 120 users. VPL can scale to a much larger set of diverse users.

Figure 4: Comparison of VPL's scalability on different tasks and user bases.

References

Ryan Boldi, Li Ding, Lee Spector, and Scott Niekum. Pareto-optimal learning from preferences with hidden context, 2024. URL https://arxiv.org/abs/2406.15599.

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, 14643510. URL http://www.jstor.org/stable/2334029.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023. URL https://arxiv.org/abs/1706.03741.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2310.12773.

Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based rl without a reward function, 2023. URL https://arxiv.org/abs/2305.15363.

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, 2023. URL https://arxiv.org/abs/2310.11564.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback, 2023. URL https://arxiv.org/abs/2303.05453.

Cassidy Laidlaw and Stuart Russell. Uncertain decisions facilitate better preference learning, 2021. URL https://arxiv.org/abs/2106.10394.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. Scalable agent alignment via reward modeling: a research direction, 2018. URL https://arxiv.org/abs/1811.07871.

Xinyu Li, Zachary C. Lipton, and Liu Leqi. Personalized language modeling from personalized human feedback, 2024. URL https://arxiv.org/abs/2402.05133.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Sriyash Poddar, Yanming Wan, Hamish Ivision, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. 2024.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf, 2024. URL https://arxiv.org/abs/2312.08358.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

Siyan Zhao, John Dang, and Aditya Grover. Group preference optimization: Few-shot alignment of large language models, 2024. URL https://arxiv.org/abs/2310.11523.