# PRIORITY2REWARD: Incorporating Healthworker Preferences for Resource Allocation Planning

**Shresth Verma** [1], **Alayna Nguyen** [1], **Niclas Boehmer**[1,2], **Lingkai Kong** [1], **Milind Tambe** [1]
[1]Harvard University
[2]Hasso Plattner Institute, University of Potsdam
sverma@g.harvard.edu, alaynanguyen@g.harvard.edu, niclas.boehmer@hpi.de, lingkaikong@g.harvard.edu, tambe@g.harvard.edu

## Abstract

In this paper, we present PRIORITY2REWARD a Large Language Model (LLM) based application which incorporates health worker preferences for resource allocation planning in public health programs. LLMs are increasingly used to design reward functions based on human preferences in Reinforcement Learning problems. We focus on LLM-designed rewards for Restless Multi-Armed Bandits, a framework for allocating limited resources among agents. In the context of public health, our approach empowers grassroots health workers to tailor automated allocation decisions to community needs. We showcase a simulated application of PRIORITY2REWARD for a large-scale mobile health program in India. The tool allows health workers to enter natural language preferences and leverages LLMs to search for reward functions aligned with the preference. Our tool then dynamically showcases how the LLM generated reward function modifies the policy outcomes with respect to different demographic groups in the population. This can help inform policy implementation at a community level.

## Overview

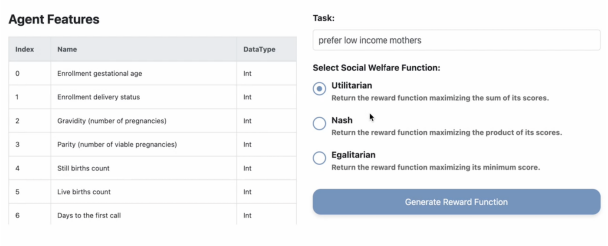
Reward functions play a fundamental role in the generation of optimal policies for sequential decision-making via reinforcement learning. Previous work has shown that LLMs are an effective tool for designing reward functions that can be guided and customized via human language prompts (Ma et al. 2023; Cao et al. 2024; Kwon et al. 2023; Xie et al. 2024; Ma et al. 2024; Yu et al. 2023; Hazra et al. 2024). We study the problem of designing high-quality reward functions aligned with human preference prompts in the context of sequential resource allocation problem in mobile health. We present a tool that proposes reward functions aligned with natural language preferences.

Specifically, the tool is designed for reward design in restless multi-armed bandits (RMABs), a popular model in multiagent systems for sequentially allocating limited resources to a set of agents (Whittle 1988; Niño-Mora 2023). In RMABs, each agent is represented by an individual Markov Decision Process including a reward function. By choosing these reward functions, one can control which agents are more or less likely to receive a resource. RMABs have been used in various domains such as machine maintenance (Abbou and Makis 2019), anti-poaching (Qian et al. 2016), and healthcare (Ayer et al. 2019; Verma et al. 2023). In many of them, system organizers have evolving allocation priorities based on agents' features that need to be incorporated into the resource allocation process (Deardorff et al. 2018; Verma et al. 2024). In a healthcare program, this translates to health care worker wishing to change the allocation policy to prioritize low-income beneficiaries who are at higher risk or older beneficiaries who have transportation barriers for healthcare access (Nelson et al. 2016; Syed, Gerber, and Sharp 2013) via the preference prompt: *Prioritize low-income beneficiaries and older beneficiaries*. PRIORITY2REWARD allows health workers to accomplish this, thus incorporating their expertise for community level planning.

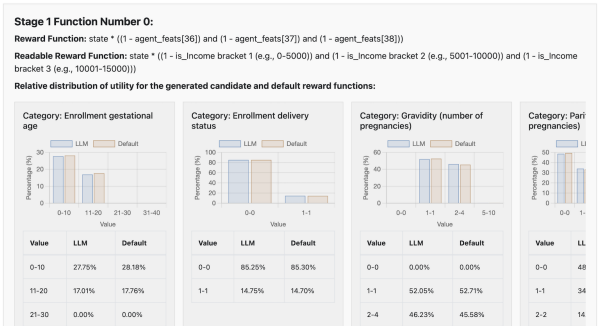


## Mobile Health Program

ARMMAN (2024) is a non-profit in India that operates large-scale Maternal and Child Care Mobile Health programs for underserved communities. One of their programs disseminates critical health information via weekly automated voice messages for expectant and pregnant mothers. The goal of the non-profit is to maximize beneficiaries' engagement, i.e., the number of messages they listen to. Given the limited number of health workers, the planning problem of scheduling these service calls (or interventions) has previously been modelled using RMAB (Mate et al. 2022). In particular, beneficiaries' listenership behaviour is modelled as a two state Markov Decision Process (MDP) where the Engaging or Non-Engaging state is defined by whether the beneficiary has listened to 30 seconds or more of an automated voice call in a week. We consider two actions for every beneficiary, the active or passive actions of making a service call or not. Finally, the planner must provide a schedule of which beneficiaries to call in a given week. Each beneficiary is also characterized by their income, education, age and other socio-demographic attributes. Beneficiaries' historic listenership values are used to estimate their transition probabilities (Mate et al. 2022). For the purpose of this demo, we use anonymized data from a quality improvement study conducted in January 2022 (Verma et al. 2023). We consider a subset of 2100 beneficiaires and a budget of 210 weekly service calls. Additionally, we consider a time horizon of 12 weeks for simulating beneficiaires' behaviour.




(a)




(b)

Figure 1: Workflow of PRIORITY2REWARD. a) The health worker enters a preference in natural language in the dashboard. b) The backend prompts an LLM to propose multiple reward functions and simulates their effect on policy outcomes and finally chosen reward functions are displayed to the health worker.

## Decision Language Model

PRIORITY2REWARD uses a Decision Language Model framework (Behari et al. 2024), in the backend to propose reward functions aligned with the human-preference. Decision Language Model consists of a generation, simulation and selection phase. The generation phase queries an LLM to propose reward function for the RMAB problem given the problem context, the available features and the health worker preference prompt. The simulation phase computes the output of the RMAB policy resulting from the modified reward function. Finally, the selection phase consists of querying an LLM to decide the best reward function which has policy output aligning most with the health worker preference.

While DLM has shown considerable benefit in designing reward functions that match those designed by human experts (Behari et al. 2024), it is computationally expensive in practice. To make the DLM propose and select reward functions in real-time, which is crucial when used by health workers, we make two modifications to DLM framework:

**Whittle Index Policy** The DLM uses PreFeRMAB (Zhao et al. 2024) as the computation method for deciding an RMAB policy. However, it is computationally expensive and we noticed that in practice, it can't scale for more than a few hundred arms in an RMAB problem. Instead, we use a heuristic known as the Whittle Index which intuitively calculates the value of active action as compared to passiver action. Whittle Index is relatively inexpensive to compute has been shown to work well in practice (Mate et al. 2022; Verma et al. 2023).

**Caching** We further improve the real-time capabilities of PRIORITY2REWARD by caching whittle indices. In particular, Whittle Indices are dependent on an arm's MDP which is characterized by transition probabilities and the reward function. We observed that often, LLM-proposed reward functions might be different in textual structure but result in the same reward for a large proportion of arms in the population. We can thus use previously computed whittle indices in such scenarios, thereby improving the speed of PRIORITY2REWARD. Together, these two improvements provide $\sim 5\times$ speedup for reward function proposal in PRIORITY2REWARD.

## Hyperparameters

We use Gemini pro (Team et al. 2023) as LLM and query it using Google Cloud API. We generate 4 reward functions for each preference prompt and conduct 10 RMAB simulations with different seeds to compute policy outcomes.

## PRIORITY2REWARD Design

DLM tool starts a workflow (see Figure 1) with a health worker writing preference in natural language in an input field. This starts the DLM loop which asynchronously updates the reward functions proposed by the LLM. In addition to reward functions, PRIORITY2REWARD also showcases how these reward function change the policy outcome. We have two views measuring the impact of reward funciton on policy outcome - i) resource distribution view and ii) reward distribution view. Resource distribution shows how the resource or in our case the service call interventions are distributed across different socio-demographic groups and beneficiary attributes. The reward distribution view, on the other hand, shows the distribution of engagement with service calls. For both of these views, we also include the distributions from a default reward function, which only maximizes engagement for everyone, for comparison. The health worker can further drill down on either of these views through interactive charts.

## Conclusion and Social Impact

We present a technology demonstration of Large Language Models for reward design to dynamically shape policy outcomes in a mobile health program. Our tool is especially useful in allowing experts at the grassroots - community volunteers and health workers to participate in decision-making by incorporating their insights into policy through natural language. This helps decentralize planning and build community-tailored solutions. In addition, the tool also allows for rapid adaptation to changing government regulations.

## Acknowledgements

This work was supported by the Harvard Data Science Initiative.

## References

Abbou, A.; and Makis, V. 2019. Group Maintenance: A Restless Bandits Approach. *INFORMS J. Comput.*, 31(4): 719–731.

ARMMAN. 2024. ARMMAN: Advancing Reduction in Mortality and Morbidity of Mothers, Children, and Neonates. https://armman.org/.

Ayer, T.; Zhang, C.; Bonifonte, A.; Spaulding, A. C.; and Chhatwal, J. 2019. Prioritizing hepatitis C treatment in US prisons. *Operations Research*, 67(3): 853–873.

Behari, N.; Zhang, E.; Zhao, Y.; Taneja, A.; Nagaraj, D.; and Tambe, M. 2024. A Decision-Language Model (DLM) for Dynamic Restless Multi-Armed Bandit Tasks in Public Health. *CoRR*, abs/2402.14807.

Cao, Y.; Zhao, H.; Cheng, Y.; Shu, T.; Liu, G.; Liang, G.; Zhao, J.; and Li, Y. 2024. Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *CoRR*, abs/2404.00282.

Deardorff, K. V.; Rubin Means, A.; Ásbjörnsdóttir, K. H.; and Walson, J. 2018. Strategies to improve treatment coverage in community-based public health programs: a systematic review of the literature. *PLoS neglected tropical diseases*, 12(2): e0006211.

Hazra, R.; Sygkounas, A.; Persson, A.; Loutfi, A.; and Martires, P. Z. D. 2024. REvolve: Reward Evolution with Large Language Models for Autonomous Driving. *CoRR*, abs/2406.01309.

Kwon, M.; Xie, S. M.; Bullard, K.; and Sadigh, D. 2023. Reward Design with Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ma, Y. J.; Liang, W.; Wang, G.; Huang, D.; Bastani, O.; Jayaraman, D.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Eureka: Human-Level Reward Design via Coding Large Language Models. *CoRR*, abs/2310.12931.

Ma, Y. J.; Liang, W.; Wang, H.; Wang, S.; Zhu, Y.; Fan, L.; Bastani, O.; and Jayaraman, D. 2024. DrEureka: Language Model Guided Sim-To-Real Transfer. *CoRR*, abs/2406.01967.

Mate, A.; Madaan, L.; Taneja, A.; Madhiwalla, N.; Verma, S.; Singh, G.; Hegde, A.; Varakantham, P.; and Tambe, M. 2022. Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12017–12025.

Nelson, L. A.; Mulvaney, S. A.; Gebretsadik, T.; Ho, Y.-X.; Johnson, K. B.; and Osborn, C. Y. 2016. Disparities in the use of a mHealth medication adherence promotion intervention for low-income adults with type 2 diabetes. *Journal of the American Medical Informatics Association*, 23(1): 12–18.

Niño-Mora, J. 2023. Markovian restless bandits and index policies: A review. *Mathematics*, 11(7): 1639.

Qian, Y.; Zhang, C.; Krishnamachari, B.; and Tambe, M. 2016. Restless Poachers: Handling Exploration-Exploitation Tradeoffs in Security Domains. In Jonker, C. M.; Marsella, S.; Thangarajah, J.; and Tuyls, K., eds., *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems, Singapore, May 9-13, 2016*, 123–131. ACM.

Syed, S. T.; Gerber, B. S.; and Sharp, L. K. 2013. Traveling towards disease: transportation barriers to health care access. *Journal of community health*, 38: 976–993.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Verma, S.; Singh, G.; Mate, A.; Verma, P.; Gorantla, S.; Madhiwalla, N.; Hegde, A.; Thakkar, D.; Jain, M.; Tambe, M.; and Taneja, A. 2023. Expanding impact of mobile health programs: SAHELI for maternal and child care. *AI Mag.*, 44(4): 363–376.

Verma, S.; Zhao, Y.; Sanket Shah, N. B.; Taneja, A.; and Tambe, M. 2024. Group Fairness in Predict-Then-Optimize Settings for Restless Bandits. openreview.net/pdf?id=GJlZbpLWX3.

Whittle, P. 1988. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A): 287–298.

Xie, T.; Zhao, S.; Wu, C. H.; Liu, Y.; Luo, Q.; Zhong, V.; Yang, Y.; and Yu, T. 2024. Text2Reward: Reward Shaping with Language Models for Reinforcement Learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

Yu, W.; Gileadi, N.; Fu, C.; Kirmani, S.; Lee, K.; Arenas, M. G.; Chiang, H. L.; Erez, T.; Hasenclever, L.; Humplik, J.; Ichter, B.; Xiao, T.; Xu, P.; Zeng, A.; Zhang, T.; Heess, N.; Sadigh, D.; Tan, J.; Tassa, Y.; and Xia, F. 2023. Language to Rewards for Robotic Skill Synthesis. In Tan, J.; Toussaint, M.; and Darvish, K., eds., *Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, 374–404. PMLR.

Zhao, Y.; Behari, N.; Hughes, E.; Zhang, E.; Nagaraj, D.; Tuyls, K.; Taneja, A.; and Tambe, M. 2024. Towards a Pre-trained Model for Restless Bandits via Multi-arm Generalization. IJCAI.