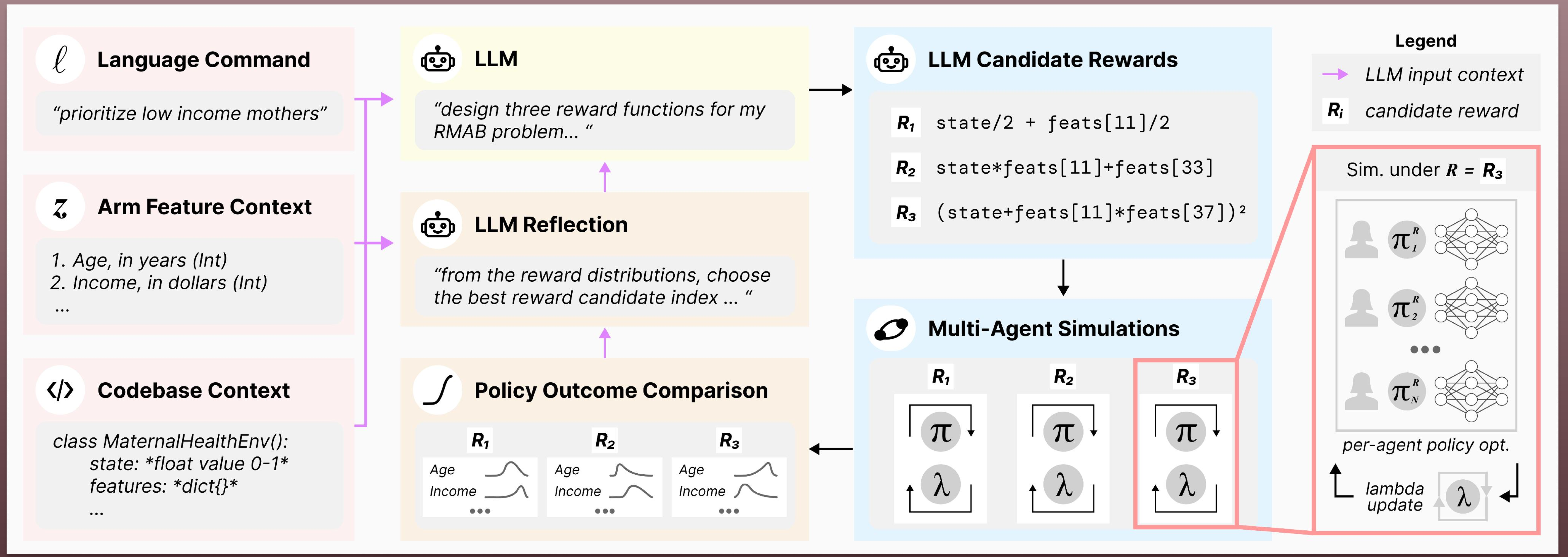
Decision Language Model: Dynamic Restless Multiarmed Bandits for Public Health



Overview of the DLM reward design loop: The LLM is provided with three inputs—a language command, a list of per-arm demographic features, and syntax cues for generating reward functions in code. Based on this, the LLM is 1) prompted to 2) generate candidate reward functions, which are used to 3) train optimal policies. These policies are simulated to produce 4) outcome comparisons across demographic groups. Finally, the LLM performs 5) self-reflection to select the best reward function, guiding future reward design.

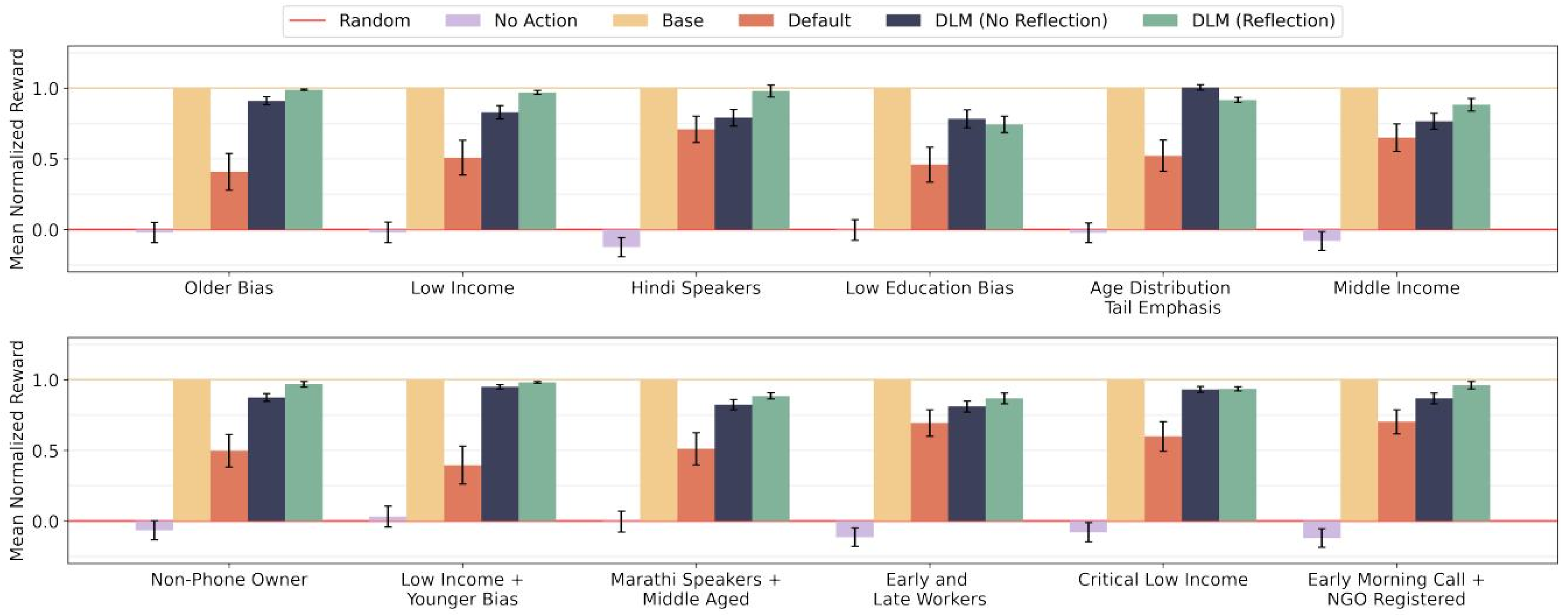
Nikhil Behari* (1,2), Edwin Zhang* (1,5,6), Yunfan Zhao (3,1), Aparna Taneja (4), Dheeraj Nagaraj (4), Milind Tambe (1,4)

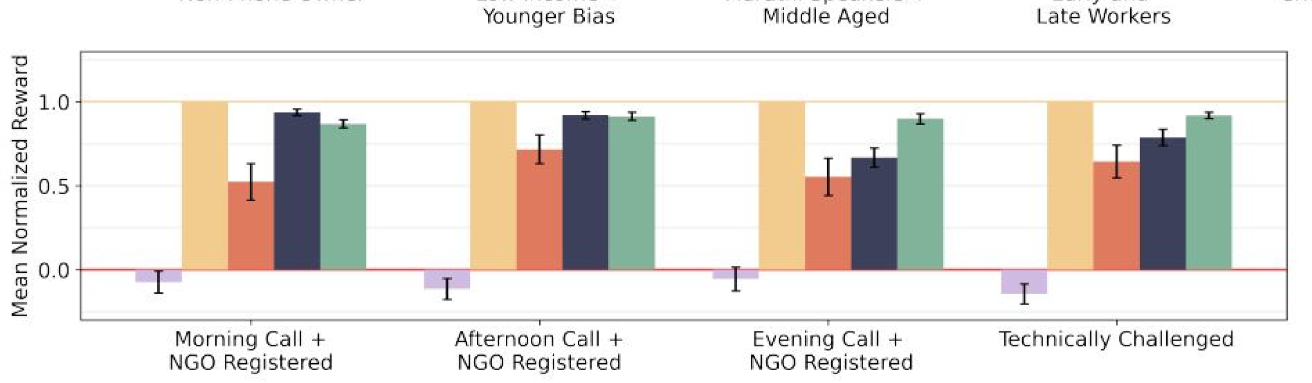
1 Harvard University 2 MIT 3 GE Healthcare 4 Google 5 OpenAI 6 Humanity Unleashed

Brief Summary

1. Restless multi-armed bandits (RMABs) have been effective in optimizing resource allocation for large public health populations but lack adaptability to changing policy priorities.
2. Large Language Models (LLMs) have proven capable as automated planners in various domains, including robotic control.
3. This paper introduces a Decision Language Model (DLM) for RMABs, allowing policy fine-tuning via human-language commands.
4. The DLM uses LLMs to interpret policy prompts, propose reward functions as code, and refine them based on RMAB simulation feedback.
5. A simulation with ARMMAN, an Indian non-profit, demonstrates DLM's ability to dynamically adjust health worker call policies using human input.

Results





	Avg. over given prompts
Base	1.00 (topline)
No Action	-0.06 ± 0.017
Default	0.57 ± 0.027
DLM (No Reflection)	0.87 ± 0.008
DLM (Reflection)	0.93 ± 0.006

Main results. We compute normalized reward (Section 5.2) for each method over 200 seeds, and report the interquartile mean (IQM) and standard error of the IQM across all runs [47]. We compare the topline Base reward policy to the performance of DLM with No Reflection and with Reflection. We also compare to a No Action and Random policy, and a Default policy that demonstrates how the original (fixed) reward function would perform for each new task. Our method is able to achieve near-base reward performance across tasks, and consistently outperform the fixed Default reward policy in a completely automated fashion. For some tasks, DLM with Reflection is also able to significantly improve upon zero-shot proposed reward.

Limitations and Future Work

We present a Decision Language Model (DLM) for resource allocation in public health, enabling language-driven policy adjustments in RMAB-based strategies. Testing in simulations highlights the need for further evaluation in local languages and real-world settings with ethical oversight. DLM ensures health experts can monitor outcomes and intervene to maintain safety and fairness, addressing data bias and ensuring beneficiary consent. Future work may explore fairness guarantees and participatory design to enhance ethical, community-driven public health policy optimization in resource-constrained environments.