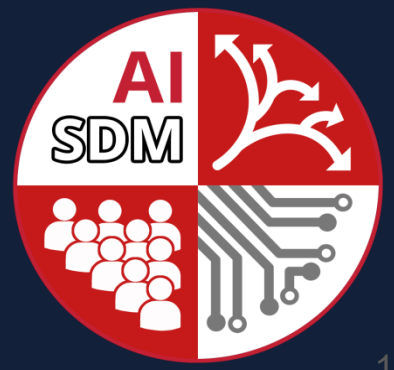
Aligning Task Utility and Human Preferences through LLM-Guided Reward Shaping

Guojun Xiong advised by Prof. Milind Tambe

Postdoctoral Fellow at Teamcore
Computer Science, SEAS, Harvard University

https://arxiv.org/pdf/2509.16399



AI for Social Impact (AI4SI) in Teamcore

- Improve decision making using AI to benefit society

Public Health

Conservation

Public Safety and Security

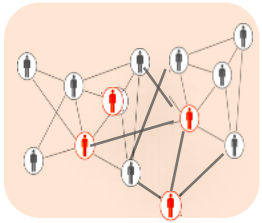
Optimize Our Limited Intervention Resources

Social Impact ⟷ *AI Innovation*

- Maximizes the utility for the entire system
- These objectives often represent years of institutional learning and proven operational success

AI Systems Align with Human Values: the Challenge

Public Health



Conservation



Public Safety & Security


Domain-Specific solvers

RMAB Solver

Green Game Solver

MDP Solver

Dynamic Preference Requirement from Humans

"Prioritize older patients more during flu season"

"Slight preference to species that are culturally significant"

"Prioritize regions with higher population density"

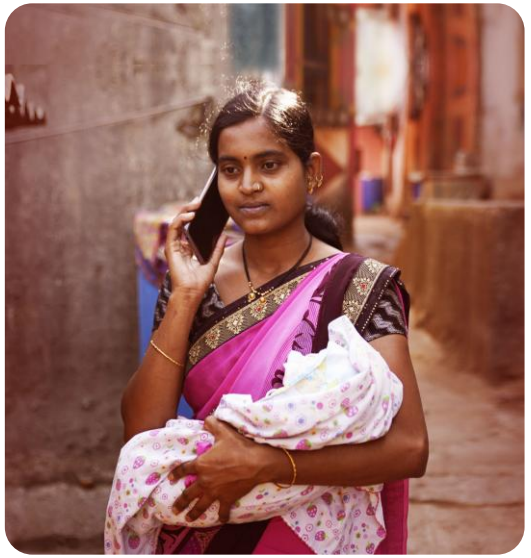
Traditional solvers cannot automatically handle additional preference requirements!

Harvard John A. Paulson School of Engineering and Applied Sciences

TEAMCORE

Example: Resource Allocation for Maternal Mothers

20	States in India
41,740,196	Beneficiaries
347,794	Workers trained
97	Partner hospitals
40	Partner NGOs



armman
(Advancing Reduction in Mortality and Morbidity of Mothers, Children and Neonates)
Delivering India's Future

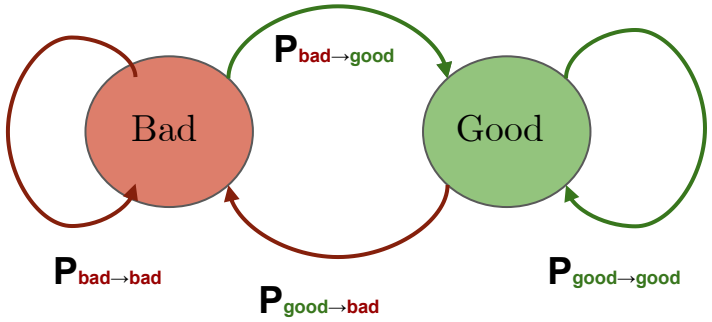
Mission: Reduce maternal, neonatal and child mortality and morbidity in underprivileged communities

Example: Resource Allocation for Maternal Mothers

Limited Resources: Service Call Allocation Problem

- Model Each of N beneficiaries as a Markov Decision Process (MDP)
- ***Find B arms to pull***
- ***Maximize the beneficiaries' engagement for the overall system***

States - Binary Valued



A "bad" state and a "good" state
s = 0 s = 1

Actions


passive

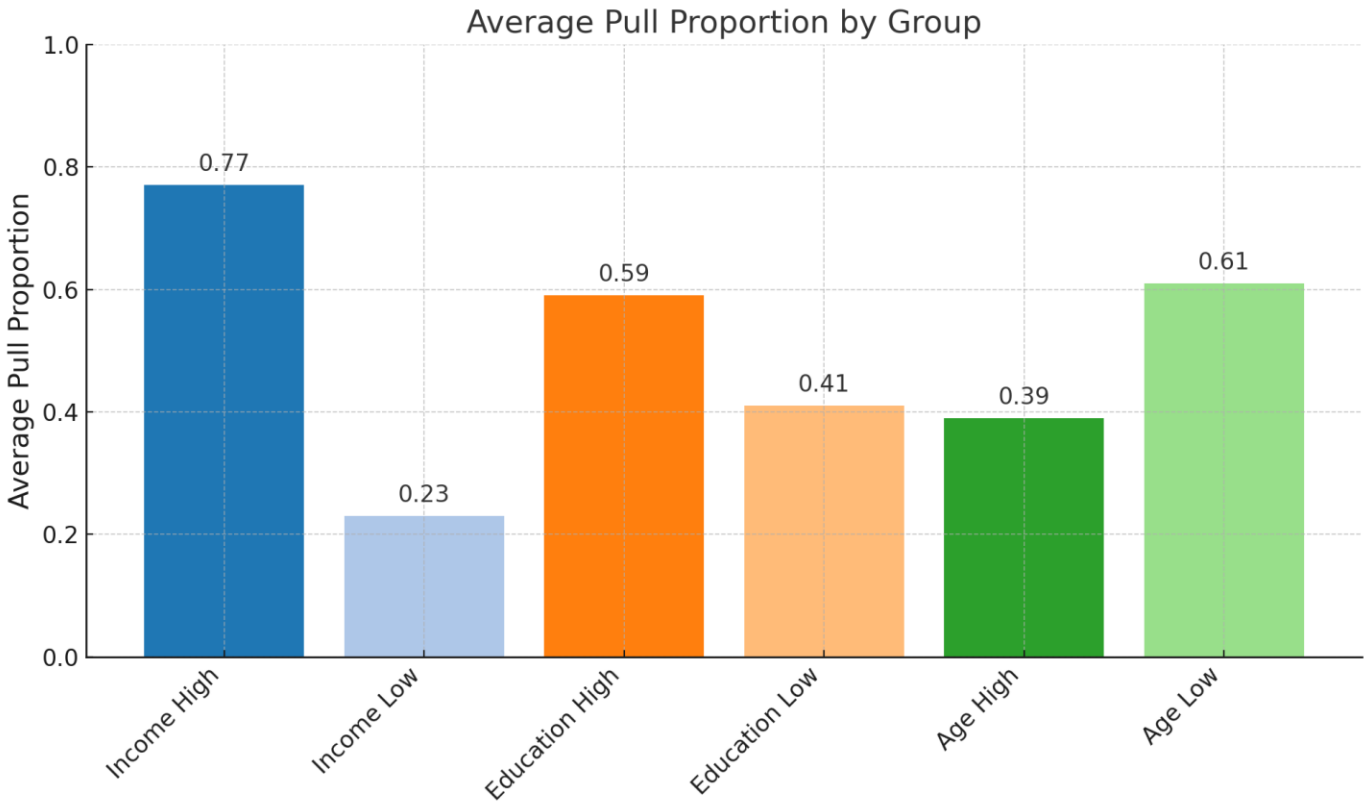

active

Place Service Call or
Not Place Service Call

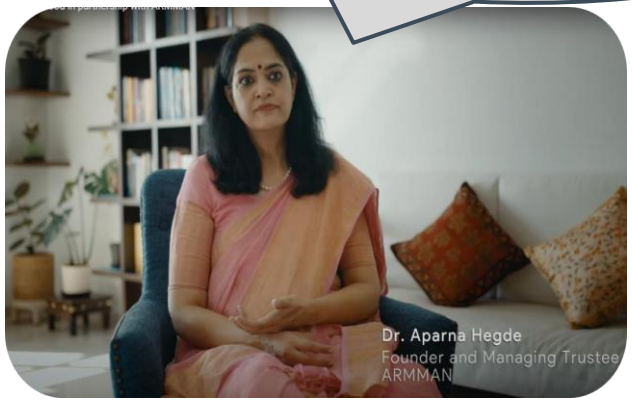
Transition matrix

[Mate. AAAI 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 2022.

Example: Resource Allocation for Maternal Mothers



Slightly Prioritize disadvantaged groups: low education, low income, and old maternal mothers



- Some groups are favored, while some groups are not
- The favored groups are the easiest to engage by inherent nature (high-income, high-edu, and young)

Traditional solvers cannot automatically handle additional preference requirements

Problem Formulation

Problem Formulation

- Consider a family of constrained sequential decision-making problems
- Define a population of N units
- Each unit is modeled an MDP
- "Resource/budget constraint": interact B out of N units at each time

Task Utility for Original System:

$$\max_{\pi \in \Pi_{feasible}} \quad U(\pi) := \mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\sum_{n=1}^{N} R_{\text{base},n}\left(S_n(t), A_n(t)\right)\right]$$

$$\text{subject to} \quad \sum_{n=1}^{N} A_n(t) \leq B, \forall t \in \mathcal{T}.$$

Maximize the total expected reward for entire system

Problem Formulation

- Each unit is represented by features capturing domain-specific attributes

[*Age: Young, Old;*
Education: High, Low;
Income: High, Low;]

- Human decision-makers often have varied soft or imprecise preference (additional)

$$\min_{\pi \in \Pi_{feasible}} \quad C(\pi) := \mathrm{Div}(D_\pi, D_{preference})$$

subject to $\quad \sum_{n=1}^{N} A_n(t) \le B, \forall t \in \mathcal{T}.$

Preference satisfaction:

$$D_\pi(z) = \frac{\#\text{ of units with feature } z \text{ being served}}{\#\text{ of total units being served}}$$

Minimize the preference deviation

Multi-Objective Problem

Task Utility for Original System $U(\pi)$:

Maximize the total expected reward for entire system

Preference satisfaction $C(\pi)$:

Minimize the preference deviation according to human's preference

Our goal (G)

$$\max_{\pi \in \Pi_{feasible}} (U(\pi), -C(\pi))$$

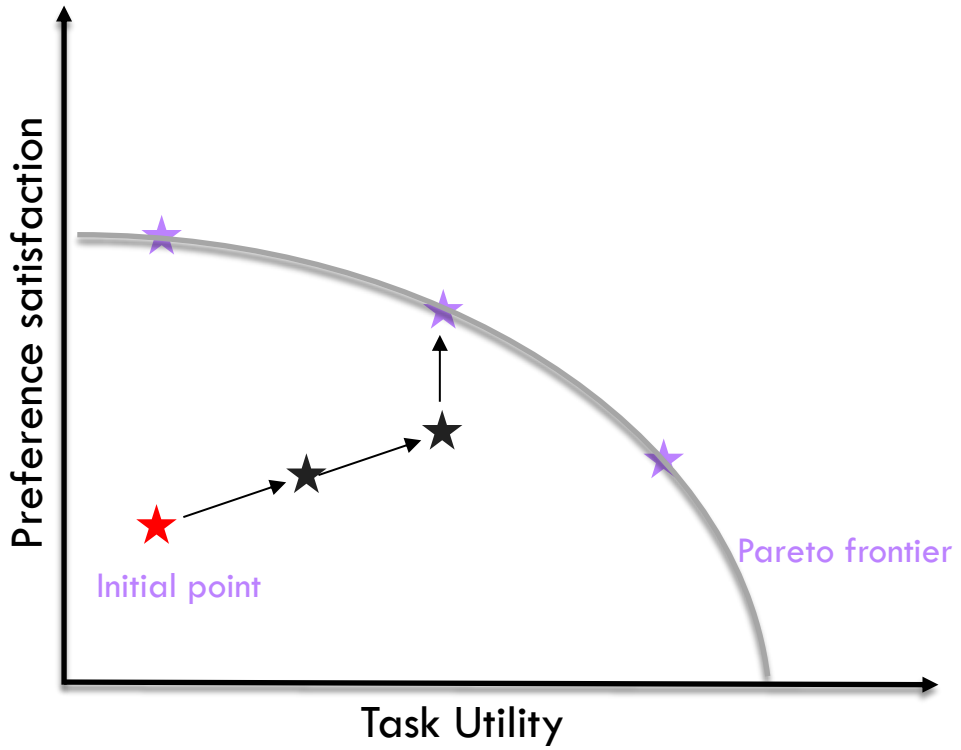
Jointly maximize the total reward and minimize the preference deviation

Pareto Frontier and Challenges of (G)

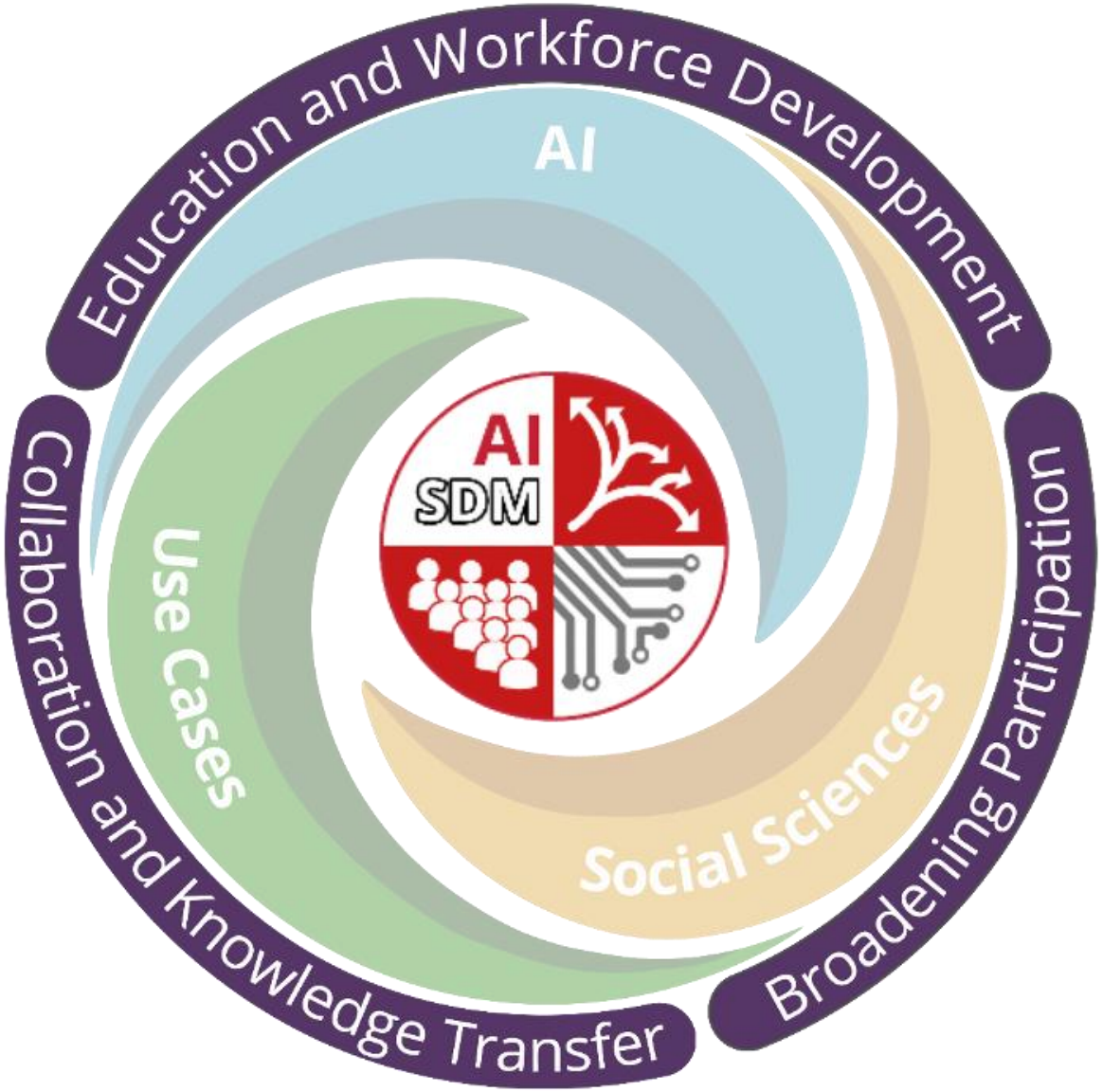
- The Pareto frontier $\mathcal{P} \subset \mathbb{R}^2$ of (G) is defined as

$$\mathcal{P} \subset \mathbb{R}^2 := \left\{ \left(U(\pi), -C(\pi)\right) \middle| \begin{array}{c} \nexists\, \pi' \in \Pi_{feasible} \; such \; that \\ U(\pi') \geq U(\pi) \\ -C(\pi') \geq -C(\pi) \end{array} \right\}$$

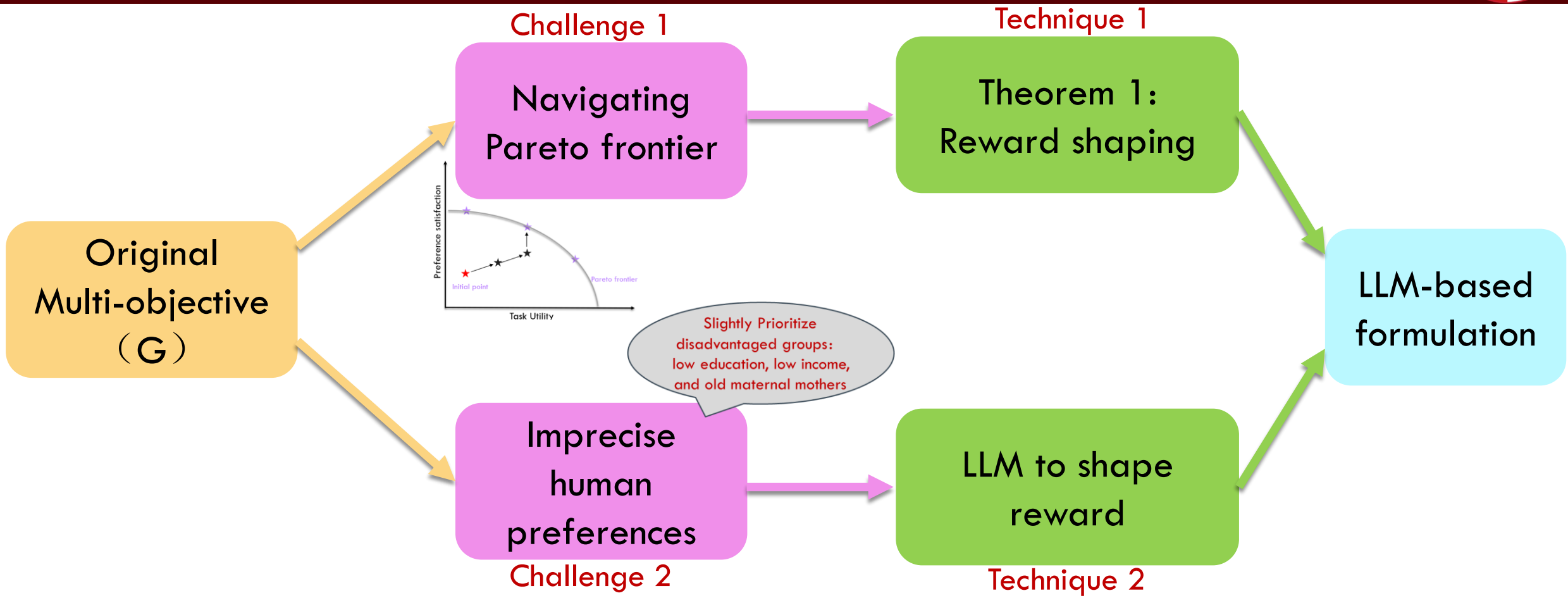
- Challenges
  - Navigating the Pareto Frontier
    - Multiple solutions on the Pareto Frontier
    - Require a precise tradeoff to balance both dimension
  - Imprecise human preferences
    - Without exact quantitative target
    - Make the divergence objective ill-defined



Proposed Method



Main Techniques of Our Method



- Key techniques
  - Reward shaping can change multi-objective into single objective
  - Leverage LLMs to shape reward

From Multi-objective to Reward Shaping

(Informal)Theorem 1: (Multi-objective to Reward Shaping). Given a predefined weight $\lambda \in [0,1]$ to balance $U(\pi)$ and $C(\pi)$, the multi-objective problem (G) is equivalent to optimizing a single objective with shaped rewards:

$$\max_{\pi} \quad J_{\lambda}(\pi) := \mathbb{E}_{\pi}\left[\sum_{t=1}^{T}\sum_{n=1}^{N} R_{shaped,n}(S_n(t), A_n(t), z_n)\right],$$

where the shaped reward is defined as:

$$R_{shaped,n}(S_n(t), A_n(t), z_n) = R_{base,n}(S_n(t), A_n(t)\ ) + R_h(\ z_n).$$

- Key takeaways
  - To solve the (G), we only need to shape the reward function by designing an additional bonus term $R_h(\ z_n)$ for the features
  - The bonus term $R_h(\ z_n)$ depends on the weight λ
  - λ is hard to be defined in practice due to the imprecise human preference (described in natural language, e.g., "slightly prefer xxx")

LLM-based Reformulation of (G)

Joint Optimization through LLM:

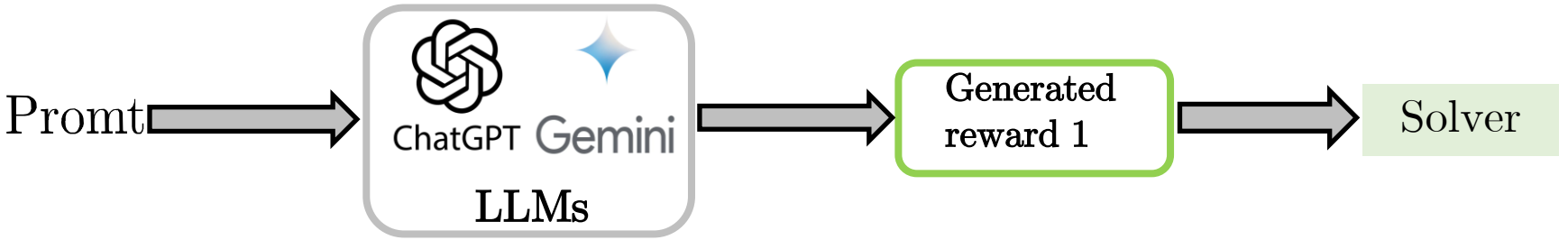
$$\max_{\text{Prompt}} \left(\underbrace{\mathrm{U}(\pi)}_{task\ utility}, \underbrace{-C(\pi, R_h)}_{Preference\ violation} \right)$$

subject to $R_h = LLM(\text{Prompt}), \pi = Solver(R_{base} + R_h)$

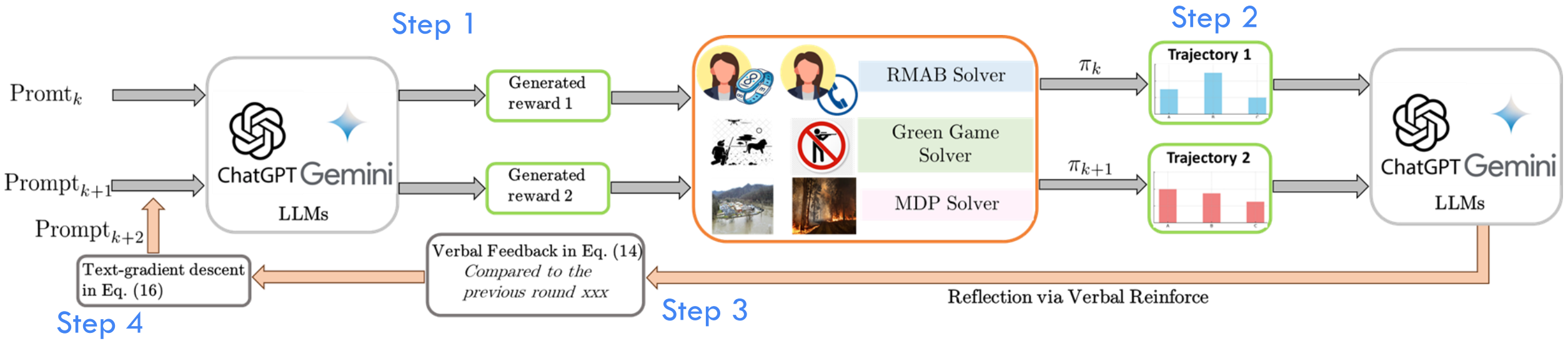
- the solver obeys all operational constraints as conventional techniques

$$\pi = \underset{\pi \in \Pi_{feasible}}{\operatorname{argmax}} \; \mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{n=1}^{N} R_{\text{base},n}\big(S_n(t), A_n(t)\big) + R_h(z_n) \right]$$

- Provided the prompt → LLM generates bonus reward → solver returns the policy



Proposed Algorithm: VORTEX



Verbal-guided Optimization with Reward Tuning via Experiential Trajectory eXploration

- Step 1: LLM-Powered Reward Generation $R_h^k = LLM(Prompt_k)$
- Step 2: Policy Execution and Evaluation (collect a trajectory for current episode k)
- Step 3: Verbal Reinforcement via Trajectoires Comparison (utility vs. preference deviation)
- Step 4: Text-Gradient Prompt Optimization (update prompt with the verbal feedback)

Exampe of Verbal Reinforcement

VORTEX Output for Public Health Domain

Starting VORTEX Algorithm (Vortex, preference: high_age)
==========
VORTEX ITERATION 1 (Vortex, prefer high_age)
==========
Querying LLM for rewards...
LLM Generated Rewards:
- age_high: +0.0200; - age_low: -0.0100; - education_high: +0.0000
- education_low: +0.0000; - income_high: +0.0000; - income_low: +0.0000

PERFORMANCE COMPARISON:
Current Utility: 8511.0 (Change: -63.0).
Current high age Coverage: 53.3% (Baseline: 47.6%).
SUCCESS: Target of 50% achieved.
RECOMMENDATION: Adjust additive feature rewards. Increase reward for 'high_age' to reach 50% coverage.

==========
VORTEX ITERATION 2 (Vortex, prefer high_age)
==========
Querying LLM for rewards...
LLM Generated Rewards:
- age_high: +0.0200; - age_low: -0.0100; - education_high: +0.0000
- education_low: +0.0000; - income_high: +0.0000; - income_low: +0.0000

PERFORMANCE COMPARISON:
Current Utility: 8496.8 (Change: -14.2).
Current high age Coverage: 54.0% (Baseline: 47.6%).
SUCCESS: Target of 50% achieved.
RECOMMENDATION: Adjust additive feature rewards. Increase reward for 'high_age' to reach 50% coverage.

- LLM reflects from the comparison and provides verbal feedback
- The verbal feedback will be feed into the prompt for next iteration reward generation

Main Theoretical Guarantee

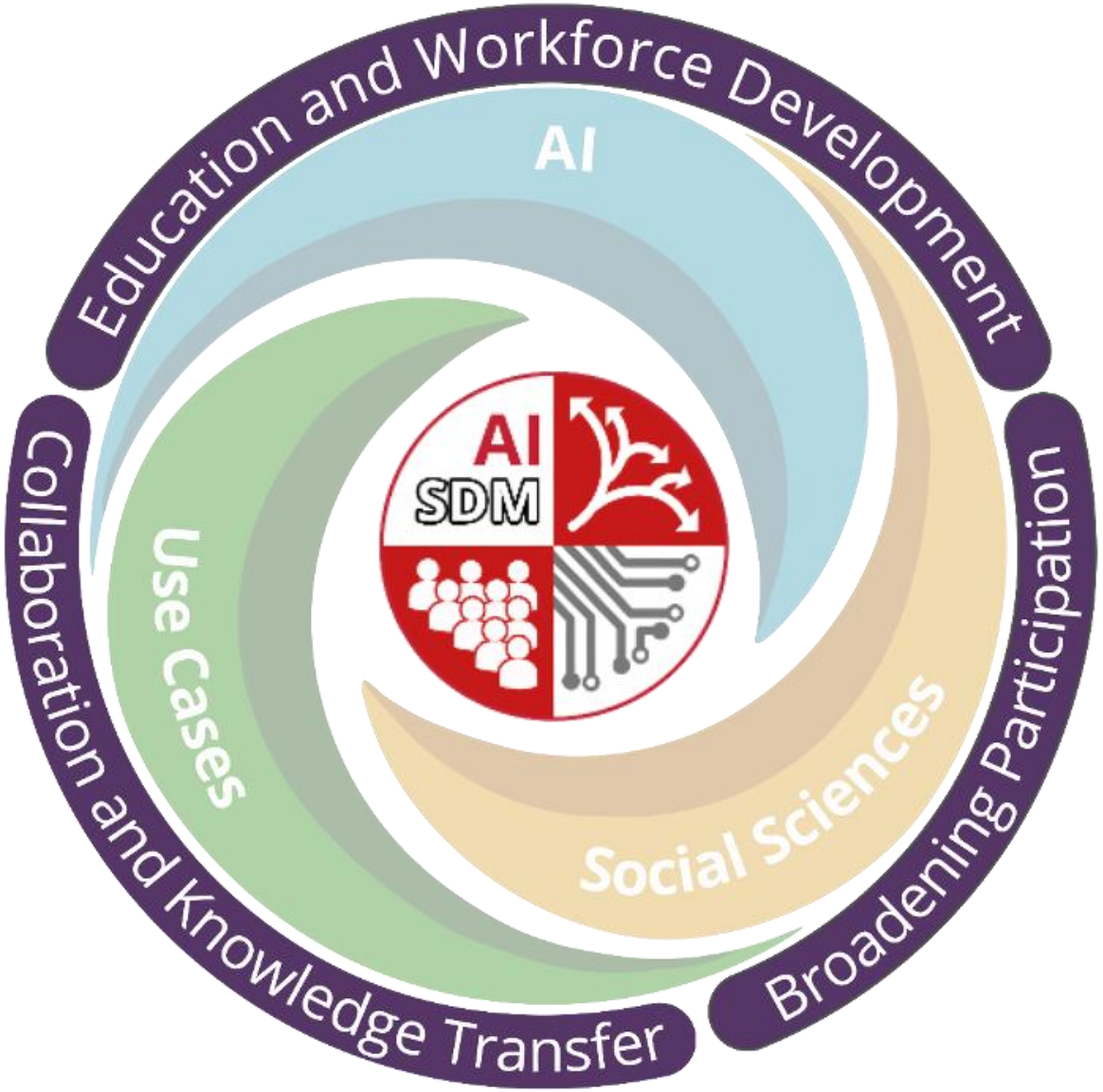
(Informal) Theorem 2 (Convergence to Pareto Optimal Point). The proposed iterative VORTEX converges almost surely to a Pareto optimal trade-off:

$$\left(U(R_h^*), C(R_h^*)\right) \in \mathcal{P}$$

- It holds when the following assumptions hold
  - External solver returns optimal policy
  - The divergence term is convex w.r.t. the feature distribution
  - The verbal reinforcement provides directional information with vanishing bias

- It provides performance guarantee of proposed VORTEX algorithm

Experiments



Experiments: ARMMAN for Maternal Health Domain

- 8 classes of mothers
- Income: Low/High; Edu: Low/High; Age: Young/Old
- Total N = 800 mothers with 100 each type
- Budget is B = 400
- State is binary $s \in \{0,1\}$
- Base reward function

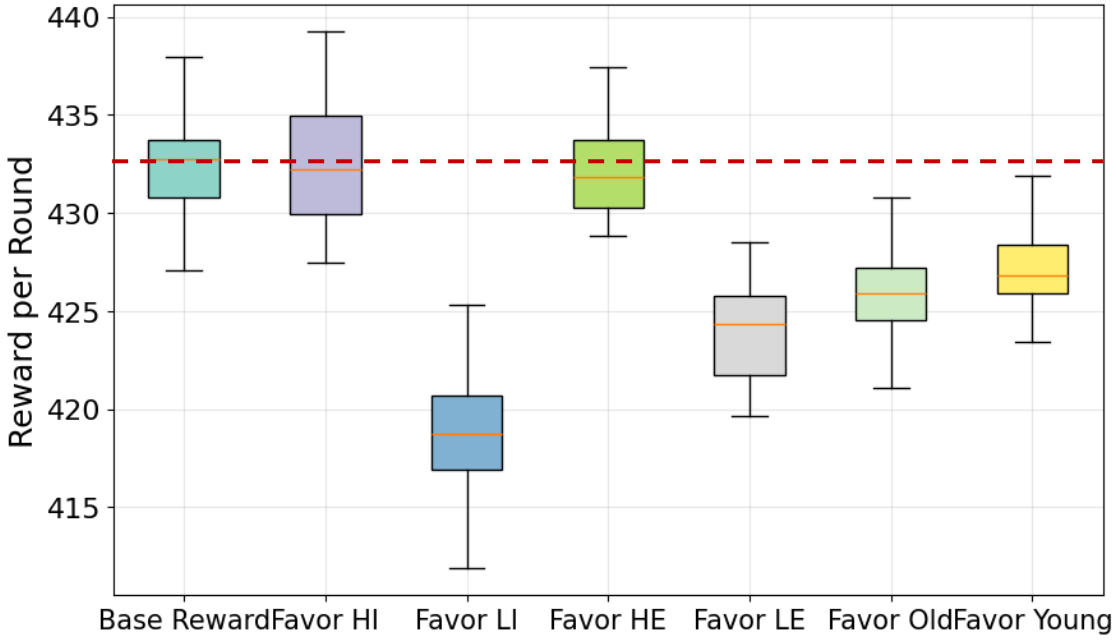
$$R_{base}(s=0) = 0.2,\ R_{base}(s=1) = 0.8$$

- 6 different preferences as: favor high/low income(HI/LI), high/low education(HE/LE), Old, and Young

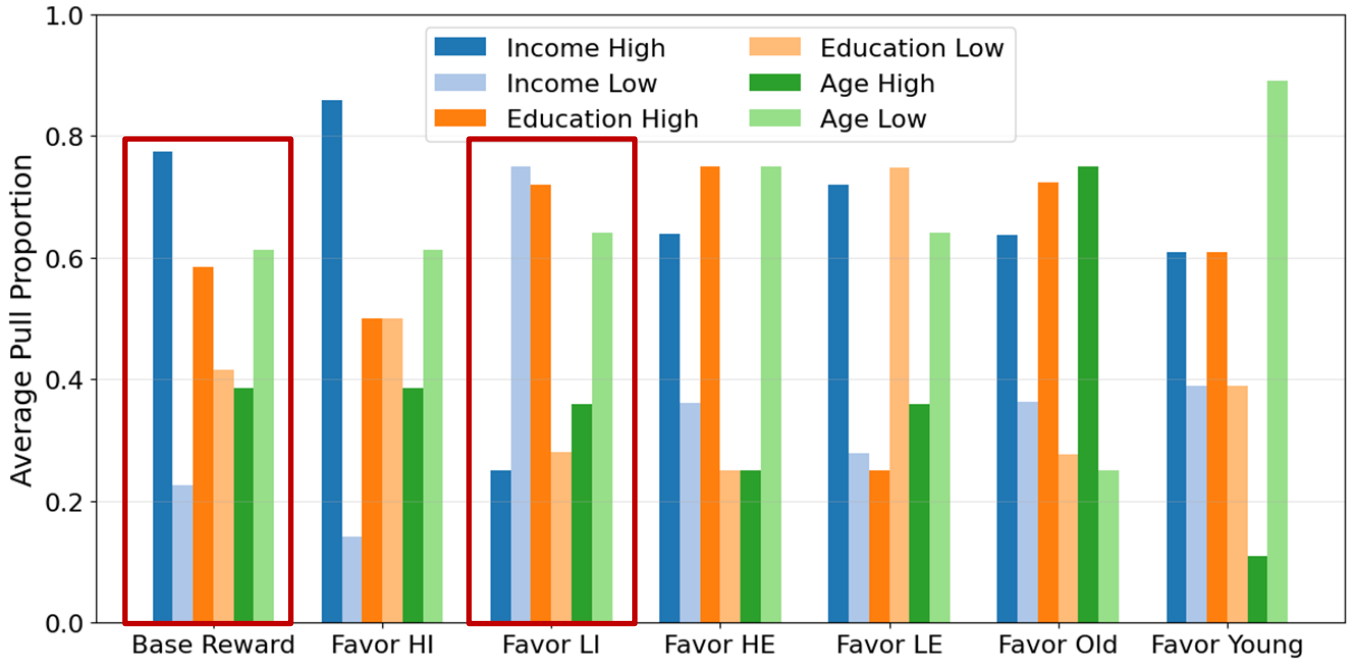
Parameter	Value
Number of patients (N)	800 pregnant women
Budget constraint (B)	400 calls per round
Time horizon (T)	50 weeks (pregnancy duration)
State space	$\mathcal{S} = \{0, 1\}$ $s = 0$: Non-adherent (high risk) $s = 1$: Adherent (low risk)

[Mate. AAAI 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 2022.

Effectiveness of Reward Shaping



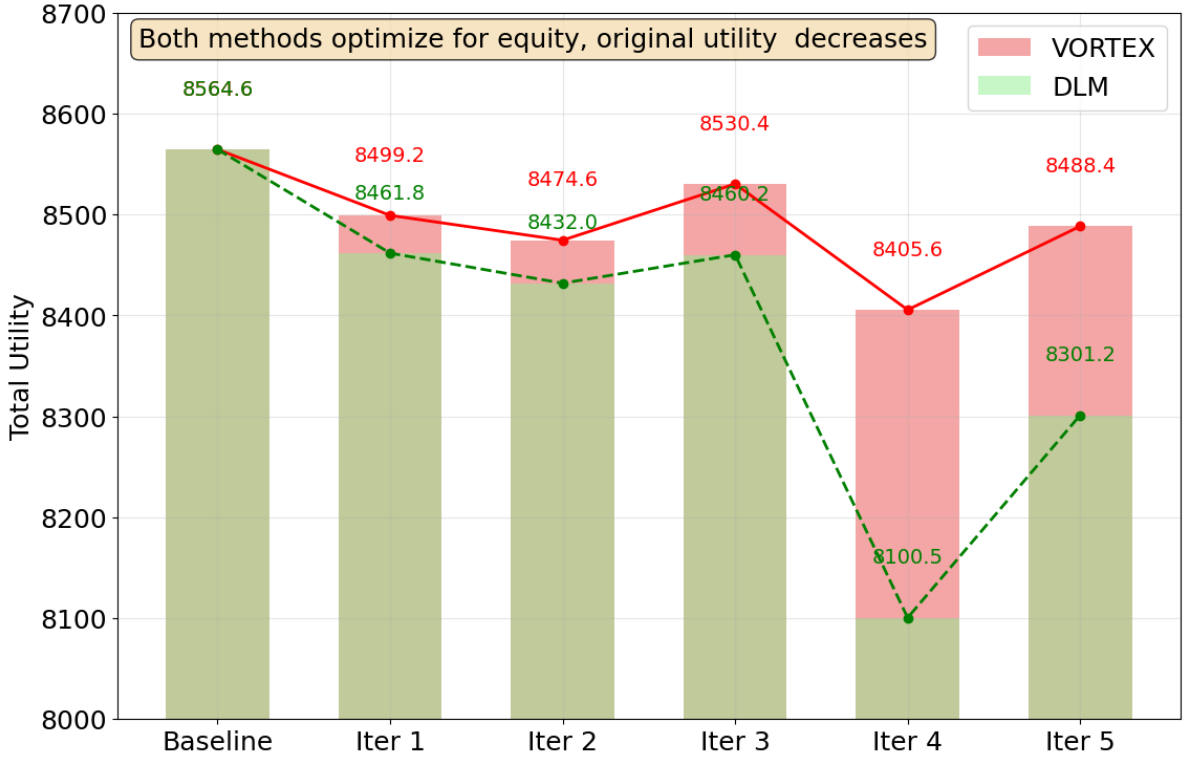
Total utility comparison



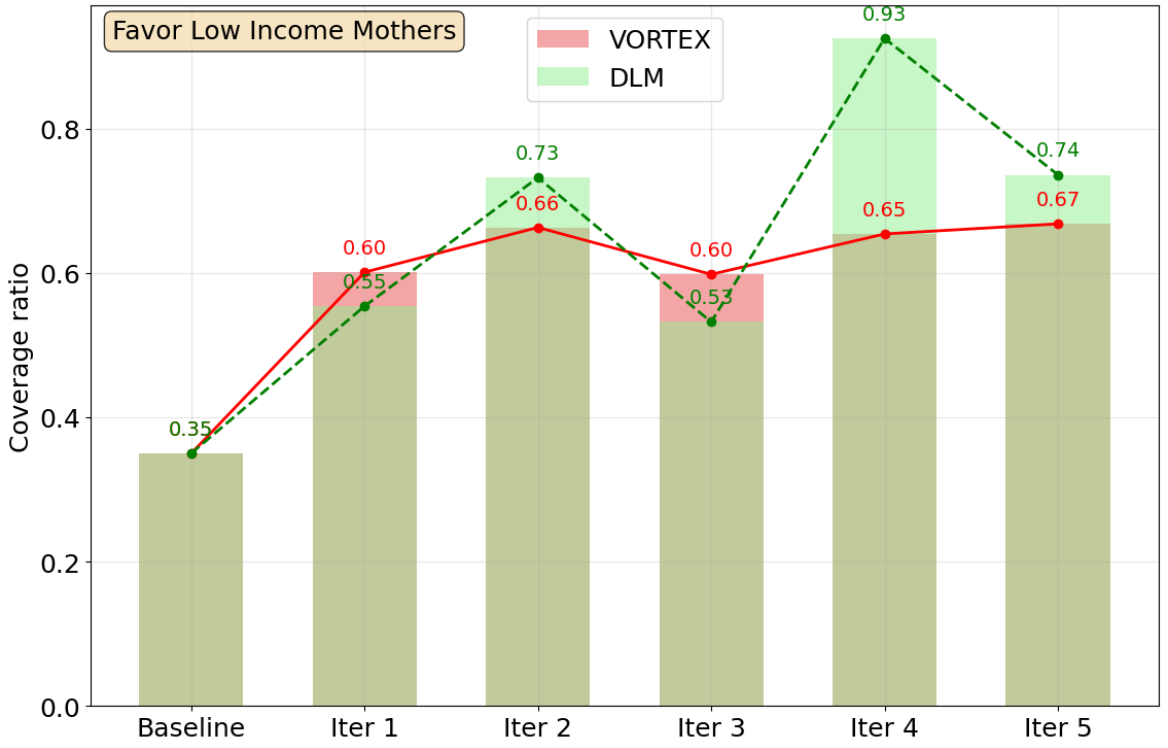
Coverage ratio comparison

- Achieve preference satisfaction with only a minimal and acceptable sacrifice in overall utility
- VORTEX can tune the coverage ratio for each type of mothers flexibly by changing the preference instructions

Baseline Comparison



Total Utility (favor low income)

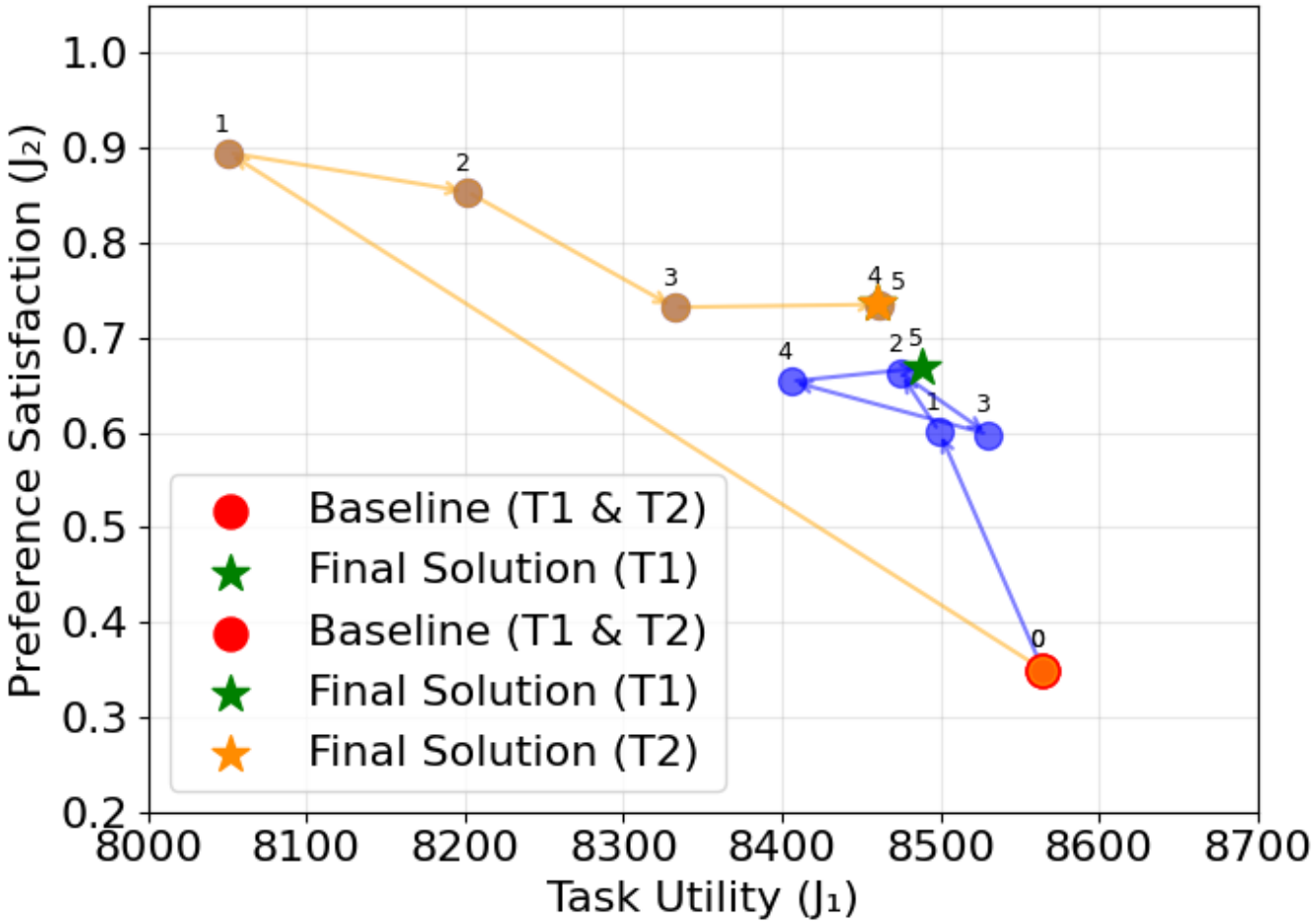


Coverage ratio (favor low income)

- Vortex is more stable and balances better on the utility and human preference than DLM

[Nikhil. Neurips2024] A decision-language model (dlm) for dynamic restless multi-armed bandit tasks in public health. In Proceedings of the Neurips2024.

Pareto Front Navigation



- Both trajectories sacrifice utility to gain preference satisfaction
- Explore different regions
- Converge to stable well-balanced solutions at different points, catering to varying stakeholder priorities

Which pareto point to select depends on the human decision-maker's preference

Conclusion

- Introduce a general multi-objective formulation to balance task untility maximization and human preference deviation minimization
  - An example of ARMMAN for Maternal Health domain
  - LLM-based reformulation
    - Prompt optimization
    - Reward shaping
  - Proposed VORTEX algorithm
    - Iterative loop
    - Low complexity
  - Numerical evaluation on Public Maternal Health Domain
    - Improved performance compared with benchmark algorithm
    - More results in other domain can be found https://arxiv.org/pdf/2509.16399

Q & A