VORTEX: Aligning Task Utility and Human Preferences Through LLM-Guided Reward Shaping

Guojun Xiong*, **Milind Tambe**

Harvard University
{gjxiong, tambe}@g.harvard.edu

Abstract

In social impact optimization, AI decision systems often rely on solvers that optimize well-calibrated mathematical objectives. However, these solvers cannot directly accommodate evolving human preferences, typically expressed in natural language rather than formal constraints. Recent approaches address this by using large language models (LLMs) to generate new reward functions from preference descriptions. While flexible, they risk sacrificing the system's core utility guarantees. In this paper, we propose VORTEX, a language-guided reward shaping framework that preserves established optimization goals while adaptively incorporating human feedback. By formalizing the problem as multi-objective optimization, we use LLMs to iteratively generate shaping rewards based on verbal reinforcement and text-gradient prompt updates. This allows stakeholders to steer decision behavior via natural language without modifying solvers or specifying trade-off weights. We provide theoretical guarantees that VORTEX converges to Pareto-optimal trade-offs between utility and preference satisfaction. Empirical results in real-world allocation tasks demonstrate that VORTEX outperforms baselines in satisfying human-aligned coverage goals while maintaining high task performance. This work introduces a practical and theoretically grounded paradigm for human-AI collaborative optimization guided by natural language.

1 Introduction

Organizations across domains deploy AI systems to optimize resource allocation with mathematical objectives carefully designed through extensive stakeholder consultation and domain expertise (Shi, Wang, and Fang 2020). For instance, public health programs allocate outreach calls to maximize patient engagement (Mate et al. 2022), conservation organizations distribute protection resources to preserve biodiversity (Dilkina et al. 2017), and emergency response systems route aid to minimize harm (Fiedrich, Gehbauer, and Rickers 2000). These objectives often represent years of institutional learning and proven operational success.

However, real-world conditions evolve rapidly. Public health crises shift demographic priorities (Bambra et al. 2020), environmental changes alter conservation needs (Pressey et al. 2007), or community feedback reveals service gaps (Abebe et al. 2020). In response, program managers frequently need to adjust their allocation strategies—perhaps to "prioritize elderly patients more during flu season" or "increase coverage for underserved rural communities." Such preference adjustments reflect operational intuition and are naturally expressed in informal language rather than mathematical formulations, especially by decision-makers who lack technical expertise to modify complex optimization functions directly (Chouldechova and Roth 2018). See Figure 1 for illustration.

his creates a fundamental tension. The existing solvers cannot directly handle the preferences and organizations are not supposed to simply abandon their carefully calibrated objectives, which embody substantial institutional knowledge and demonstrated performance. Yet ignoring evolving stakeholder preferences may lead to solutions that are mathematically optimal but operationally inadequate (Obermeyer et al. 2019). To bridge this tension, recent works (Behari et al. 2024; Verma et al. 2025) generate entirely new reward functions from natural language descriptions with the aid of large language models (LLMs). While it enables flexible preference expression and maintains compatibility with existing solvers, it provides no guarantees about preserving performance on the original optimization criteria that organizations invested significant resources to develop.

This prompts us a question: how can we achieve mathematical optimality while respecting stakeholders' preferences?

To address this, we formulate a multi-objective optimization problem that preserves the established mathematical objective while incorporating human preferences as a second, alignment-oriented objective. A common technique is weighted scalarization (Miettinen 2012; Roijers et al. 2013), which converts the problem into a single objective, and remains compatible with unmodified task-specific solvers. However, there remain two fundamental challenges. First, scalarization aims in navigating the Pareto frontier, which require users to specify explicit scalarization weights or engage in an iterative process of tuning these weights to explore the trade-off space, a significant burden for decision-makers who express preferences qualitatively. Second, these human preferences are often imprecise and expressed in natural language, making them difficult to encode as a formal mathematical objective (Christiano et al. 2017).

*Correspondence to Guojun Xiong <gjxiong@g.harvard.edu>.

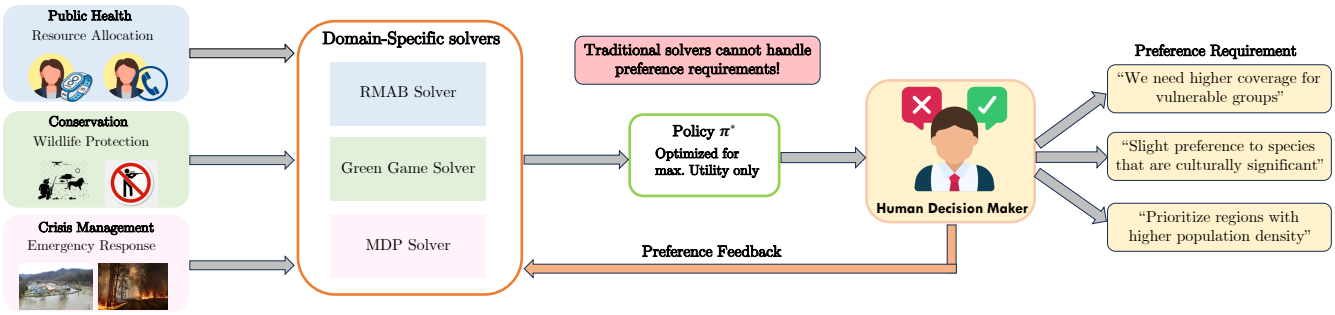


Figure 1: The fundamental tension between mathematical objective and human preferences.

To overcome these intertwined challenges, we propose VORTEX (Verbal-guided Optimization with Reward Tuning via Experiential Trajectory eXploration). To tackle the second challenge-encoding imprecise natural language preferences -VORTEX employs an LLM to generate auxiliary shaping rewards directly from this qualitative feedback. Crucially, to overcome the first challenge-navigating the Pareto frontier without the burden of manual weight tuning-VORTEX introduces an iterative framework that refines these shaping rewards through verbal reinforcement and text-gradient updates(Yu et al. 2024). Rather than replacing carefully engineered objectives, our approach augments them with adaptive, preference-aligned signals, allowing stakeholders to steer decision behavior while using existing solvers unmodified. We further provide theoretical guarantees that this process converges to Pareto-optimal solutions, achieving a principled trade-off between task utility and preference satisfaction.

Our Contributions. We summarize our main contributions as follows. ▷ *New Problem Formulation*: We cast the challenge of aligning algorithmic decision-making with human preferences as a multi-objective optimization problem, preserving core task utility while introducing a separate preference satisfaction objective. ▷*Solver-Compatible Reward Shaping*: We propose a novel framework that leverages LLMs to encode human preference objectives via reward shaping, enabling compatibility with unmodified, reward-driven solvers. ▷ *Interactive Preference Refinement:* We propose an iterative optimization procedure, VORTEX, that refines LLM prompts using verbal feedback and text-gradient updates to navigate the Pareto frontier between task performance and preference alignment without specifying tradeoff weights. ▷ *Theoretical and Empirical Validation:* We prove that VORTEX converges to Pareto-optimal solutions under mild assumptions, and we demonstrate its effectiveness on real-world resource allocation tasks, achieving improved preference satisfaction while maintaining high task utility.

2 Related Work

Reward Shaping and Preference Constraints. Reward shaping is a long-standing RL technique (Ng, Harada, and Russell 1999) and has been applied to fairness (Jabbari et al. 2017), diversity (Celis et al. 2019), and coverage (Venkataraman et al. 2021). Our work differs by generating shaping terms directly from free-form natural language via LLMs, enabling soft and imprecise constraints without explicit formalization.

LLMs for Decision Optimization. LLMs have been used to design reward functions from language descriptions (Ma et al. 2023; Li et al. 2024; Yu et al. 2023; Kwon et al. 2023). In public-health resource allocation, the DLM framework (Behari et al. 2024; Verma et al. 2025) and Kim et al. (2025b) apply LLMs to construct or fine-tune reward models that encode human preferences. These approaches, however, often fully adopt LLM-generated rewards and may override task utility. Our method instead augments a domain solver's base reward with language-derived shaping, maintaining core utility while improving preference alignment. Table 1 summarizes key differences from DLM.

Multi-objective and Constrained RL. Multi-objective RL (Roijers et al. 2013) and CMDPs (Altman 1999) traditionally rely on weighted scalarization (Miettinen 2012; Hayes et al. 2022). Exploring the Pareto frontier often requires manual weight tuning. Kim et al. (2025a) address this via p-mean welfare portfolios spanning fairness preferences. In contrast, VORTEX uses an LLM to translate verbal critiques into shaping adjustments, providing an intuitive and language-driven mechanism for navigating trade-offs.

Human-in-the-Loop Optimization. Classic approaches such as inverse RL (Ng and Russell 2000; Abbeel and Ng 2004) and preference-based RL (Christiano et al. 2017) rely on structured feedback or comparison queries and are costly to scale. We instead use LLMs as flexible oracles that directly encode human preferences into shaping rewards, enabling rapid and low-friction integration without policy retraining.

3 Problem Formulation

3.1 Problem Statement

We consider a family of decision-making problems that can be formulated as constrained stochastic optimization, such as public health resource allocation or conservation planning. Formally, we define a population of N units (e.g., patients, species, or locations), each with a state $s_i(t) \in \mathcal{S}$, and an action $a_i(t) \in \{0, 1\}$ at time t. Specifically, action $a_i(t) = 1$ represents the unit i is being allocated with a resource at time step t and action 0 otherwise. Let $P_i(s'|s, a)$ be the transition probability from state s to s' under action a for unit i. At each decision time slot, the decision maker can choose up to

Method	Multiple Objective?	Entire Reward Generation?	Theoretical Guarantee?	Tradeoff Control?
Decision Language Model				
DLM (Behari et al. 2024)	✗	✓	✗	✗
SCLM (Verma et al. 2025)				
Conventional RL				
Eureka (Ma et al. 2023)	✗	✓	✗	✗
Auto-MC (Li et al. 2024)				
VORTEX (proposed)	✓	✗	✓	✓

Table 1: We highlight the difference between our proposed method VORTEX and prior approaches. VORTEX uniquely supports multi-objective optimization, provides theoretical guarantees, and allows trade-off control—while prior methods either ignore base objectives, lack formal analysis, or offer limited preference alignment.

B units to interact, which leads to a global *budget constraint* restricts total interactions: $\sum_{i=1}^{N} a_i(t) \leq B, \quad \forall t.$

Task Utility. Let $R_{\text{base},i}(t)$ denote the task-defined reward for unit i at time t, which depends on the current state $s_i(t)$ and action $a_i(t)$, reflecting the intrinsic domain goal (e.g., a patient becoming adherent or a habitat improving). The cumulative task utility under any policy π is:

$$U(\pi) = \mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{i=1}^{N} R_{\text{base},i}(t) \right]. \tag{1}$$

Hence, the goal is to find a policy π mapping states to actions such that it maximizes the expected cumulative reward under the given dynamics:

$$\max_{\pi \in \Pi_{feasible}} U(\pi), \tag{2}$$

where $\Pi_{feasible}$ is the policy set that meets the interaction budget constraint $\sum_{i=1}^{N} a_i(t) \leq B, \quad \forall t..$

Preference satisfaction. Notice that each unit in the population is represented by a feature vector $z_i \in \mathcal{Z}$ capturing domain-specific attributes relevant to decision-making. In practice, human decision-makers often have additional soft or imprecise preference constraints that express high-level societal or ethical priorities, going beyond the base task utility maximization problem in (2). We provide a toy example to illustrate it as follows.

Example 3.1. *Consider a public health planner allocating outreach calls to a population of pregnant women (Mate et al. 2022). Each woman is described by a feature vector z_i consisting of: **Age**, **Education level**, and **Income level**, with each containing three levels, i.e., Low, Medium, High. For a feature vector $z_i = [100100100]$, it represents a "young low-education low-income" patient. Human decision-makers might require:*

"`Prioritize elderly patients slightly more than younger patients, even if it reduces short-term utility marginally.`"

The soft preference constraints are encoded as functions over the empirical feature visitation distribution for policy π:

$$D_\pi(z) = \frac{\#\text{ of units with feature } z \text{ being served}}{\#\text{ of total units being served}}. \tag{3}$$

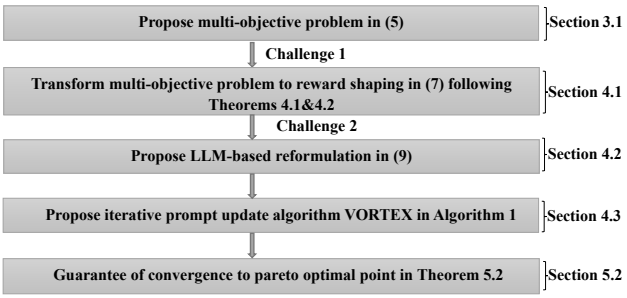


Figure 2: The main flow of contribution in this work.

Hence, it adds to a second-dimensional objective, i.e., achieving the desired demographic distribution

$$\min_{\pi \in \Pi_{feasible}} C(\pi) := \text{Div}(D_\pi, D_{\text{preference}}), \tag{4}$$

where $D_{\text{preference}}$ is the soft and imprecise preference constraint from human decision-makers, and Div is a general f-divergence measure.

Multi-objective Problem. Our goal is to find a policy π that simultaneously maximizes task utility $U(\pi)$ and minimizes preference deviation $C(\pi)$, with respect to the budget constraint $\sum_i a_i(t) \leq B$ for each t. This defines a multi-objective constrained optimization problem:

$$\max_{\pi \in \Pi_{feasible}} (U(\pi), -C(\pi)). \tag{5}$$

Pareto Frontier. The *Pareto frontier* $\mathcal{P} \subset \mathbb{R}^2$ of (5) is defined as:

$$\mathcal{P} := \left\{ (U(\pi), -C(\pi)) \,\middle|\, \begin{array}{l} \nexists\, \pi' \in \Pi_{\text{feasible}} \text{ such that:} \\ U(\pi') \geq U(\pi), \\ -C(\pi') \geq -C(\pi) \end{array} \right\}.$$

This set characterizes policies for which no other feasible solution simultaneously improves task utility and reduces preference violation. The set of all Pareto optimal solutions forms the Pareto frontier, representing all possible trade-offs between task utility and preference satisfaction.

Remark 3.2. *Modeling human preferences as a separate optimization objective, rather than a hard constraint, offers three key advantages: it avoids feasibility issues from overly strict or conflicting constraints, enables transparent trade-offs between competing goals, and allows solution quality to be evaluated via Pareto optimality (Roijers et al. 2013).*

3.2 Challenges for (5)

Challenge 1: Navigating the Pareto Frontier. While weighted scalarization (Miettinen 2012) allows us to combine task utility and preference satisfaction into a single objective, it requires a precise scalarization weight specified to navigate the resulting Pareto frontier to find a solution that aligns with stakeholder preferences. This creates a significant practical barrier, often leading to a tedious trial-and-error process of tuning weights to explore the trade-off space. To bridge this

gap, we draw on scalarization theory to transform preference alignment into a reward-shaping task. This allows us to encode human preferences as an auxiliary reward signal without tuning the weight. This full method is described Section 4.1.

Challenge 2: Imprecise human preferences. Human-specified preferences are typically high-level, or underspecified, without providing an exact quantitative target distribution. This makes the divergence objective (4) ill-defined. To address this, we propose to treat human preferences as latent objectives that can be implicitly captured through natural language by LLMs. In this way, the divergence term is approximated through LLM-based reward shaping and feedback, allowing us to integrate soft constraints into existing solvers without requiring formal definitions. This will be elaborated in Section 4.2. The main flow of contribution in this work is summarized in Figure 2.

4 Proposed Method: VORTEX

We begin by showing that the multi-objective problem in (5) can be equivalently reformulated as a reward shaping problem. We then recast it as a prompt optimization problem over LLM-generated shaping rewards. Finally, we present the full VORTEX algorithm and its iterative optimization procedure.

4.1 From Multi-objective to Reward Shaping

Provided the challenge that existing solvers fail for the multi-objective problem in (5), we leverage the scalarization theory (Miettinen 2012) of multi-objective optimization to translate (5) into a reward shaping problem. Therefore, it maintains compatibility with existing solvers.

Scalarization Theory We establish the theoretical foundation for converting multi-objective optimization to reward shaping through the following theorem.

Theorem 4.1 (Weighted Scalarization(Miettinen 2012))**.** *For the multi-objective problem defined in* (5)*, a policy π^* is Pareto optimal if and only if there exists a weight $\lambda \in [0, 1]$ such that π^* is an optimal solution to the scalarized problem:*

$$\max_{\pi} J_\lambda(\pi) = \lambda U(\pi) - (1-\lambda) C(\pi). \tag{6}$$

We now show how the scalarized objective can be reformulated as a reward shaping problem in the following theorem.

Theorem 4.2 (Multi-objective to Reward Shaping)**.** *The scalarized multi-objective problem is equivalent to optimizing a single objective with shaped rewards:*

$$\max_{\pi} J_\lambda(\pi) = \mathbb{E}_\pi \left[\sum_{i=1}^{N} \sum_{t=1}^{T} R_{\textit{shaped},i}(s_i(t), a_i(t), z_i) \right], \tag{7}$$

where the shaped reward is defined as:

$$R_{\textit{shaped},i}(s, a, z_i) = R_{\textit{base},i}(s, a) + R_h(z_i), \tag{8}$$

with shaping reward being $R_h(z) \propto \frac{1-\lambda}{\lambda} \cdot \frac{\partial}{\partial D_\pi(z)} \textit{Div}(D_\pi \| D_{\textit{preference}})$.

Corollary 4.3 (Parameter Interpretation)**.** *The scalarization parameter λ has the following interpretation: 1) $\lambda \to 1$: focus primarily on task utility; 2) $\lambda \to 0$: focus primarily on preference satisfaction. In essence, λ serves as a trade-off weight balancing these two objectives.*

This theoretical framework establishes that any Pareto optimal solution to the multi-objective problem can be found by solving a single-objective problem with appropriately shaped rewards, providing the mathematical foundation for our approach in Section 4.2.

4.2 LLM-based Reformulation of (5)

According to Theorem 4.2, the multi-objective optimization problem can be addressed through appropriate reward shaping. To tackle the second challenge, we therefore propose a novel framework that enables LLMs to generate an auxiliary reward term R_h, shifting the focus to prompt optimization for producing high-quality, preference-aligned shaping rewards.

Joint Optimization Objective. Let `Prompt` denote the input to the LLM, consisting of a fixed task description and an adaptive feedback component that evolves over time. Given `Prompt`, the LLM generates a shaping reward vector R_h, which is combined with the base reward and passed to a solver to produce a policy π. As a result, the original multi-objective problem in (5) can be reformulated as:

$$\max_{\texttt{Prompt}} \Big(\underbrace{U(\pi)}_{\text{task utility}}, - \underbrace{C(\pi)}_{\text{preference violation}} \Big)$$
$$\text{s.t. } R_h = \text{LLM}(\texttt{Prompt}), \pi = \text{Solver}(R_{\text{base}} + R_h). \tag{9}$$

Here, $C(\pi)$ measures the divergence between the state-feature visitation distribution induced by policy π and the desired preference pattern (e.g., demographic fairness), and the solver obeys all operational constraints, such as budget. The details of the solver are provided below.

Solver and Reward Composition. Given a reward function combining base and shaped components, the solver optimizes the policy under feasibility constraints:

$$\pi = \arg \max_{\pi \in \Pi_{\text{feasible}}} \mathbb{E}_\pi \left[\sum_{t=1}^{T} \sum_{i=1}^{N} (R_{\text{base},i}(t) + R_h(z_i)) \right]. \tag{10}$$

Pareto Frontier Navigation Via Language Models. Rather than manually tuning scalarization weights (e.g., Lagrange multipliers), our framework leverages LLMs to shape rewards through iterative natural language reflection. In the formulation (9), the LLM serves as a bridge between human intent and formal optimization, translating imprecise preferences into a shaped reward R_h that complements the base objective. By iteratively refining R_h based on observed trade-offs between utility U and preference violation C, the framework progressively steers the policy toward better alignment along the Pareto frontier. See Figure 7 for an illustration.

4.3 VORTEX: Description

Our proposed method, VORTEX, is an iterative algorithm that refines the LLM-generated shaping reward, R_h, through

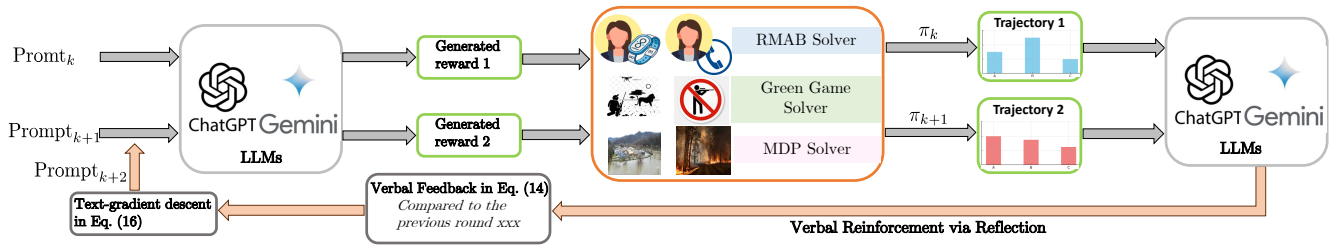


Figure 3: The detailed procedure of VORTEX. At each iteration, VORTEX compares two consecutive policy trajectories and reflects on their differences via verbal reinforcement. The resulting feedback is used to perform a text-gradient update on the LLM prompt, progressively refining the shaping reward to improve alignment.

a closed loop of generation, execution, and reflection. The goal is to progressively steer the system's policy toward a desirable point on the Pareto frontier of task utility and preference satisfaction. The core of VORTEX is an iterative process where each step builds upon the last. At each episode k, the algorithm executes four main steps: (1) Reward Generation, (2) Policy Execution and Evaluation, (3) Verbal Reinforcement via Reflection, and (4) Text-Gradient Prompt Optimization. This entire workflow is illustrated in Figure 3.

Step 1: LLM-Powered Reward Generation. Each episode begins by constructing a prompt, $\texttt{Prompt}_k$ as

> **Task:** Generate a shaping reward vector to encourage slightly higher coverage for demographic group X while preserving high total reward.
> **Context:** Each arm has state s_i and features z_i.
> **Reflection:** Previous round achieved xx% of max reward but only xx% coverage for group G.
> **Instruction:** Assign additional reward values to each arm to improve group G coverage with minimal reward loss.
> **Output:** A reward function R_h representing the preference-aligned shaping reward.

The LLM processes this comprehensive prompt to generate a new shaping reward,

$$R_h^k = \text{LLM}(\texttt{Prompt}_k). \tag{11}$$

Step 2: Policy Execution and Evaluation. The generated shaping reward, R_h^k, is added to the base task reward, R_{base}. This combined reward function is then passed to the pre-existing, unmodified domain-specific solver. The solver, operating under its constraints (e.g., budget B), calculates the optimal policy π_k. This policy is then deployed in the environment to collect one trajectory,

$$\tau^k = \left\{(s_i(t), a_i(t), z_i, R_{\text{base},i}(t)), \forall i\right\}_{t=1}^{T}.$$

Let the previous trajectory be τ^{k-1}, obtained from policy π^{k-1}. For both trajectories, we compute the total expected reward

$$U_k = \mathbb{E}_{\pi^k}\left[\sum_{t=1}^{T}\sum_{i} R_{\text{base},i}(t)\right], \tag{12}$$

and empirical feature distribution

$$D_k(z) = \frac{\#\text{ of unit with feature } z \text{ being served}}{\#\text{ of total units being served}}. \tag{13}$$

Step 3: Verbal Reinforcement via Trajectory Comparison. This step serves as the reflective engine of VORTEX. It synthesizes performance changes into structured, actionable feedback. The system compares the metrics from the current trajectory τ^k with the previous one τ^{k-1} by computing changes across iterations:

$$\delta_U = U_k - U_{k-1}, \qquad \delta_D = D_k(z) - D_{k-1}(z). \tag{14}$$

This quantitative comparison is then translated into qualitative, natural language feedback, which we term Verbal Reinforcement. This can be handled by a function or a separate LLM call,

$$\text{Feedback}_{\text{verbal}}^k = f\left(\delta_U, \delta_D\right). \tag{15}$$

The feedback explicitly states the trade-off that was observed and provides concrete suggestions for the next iteration.

Step 4: Text-Gradient Prompt Optimization In classical reinforcement learning, policies are updated via gradient descent $\theta \leftarrow \theta - \eta\nabla_\theta J(\theta)$. In our framework, we operate in the space of prompts rather than parameters. We decompose the prompt into two disjoint components:

$$\texttt{Prompt}_k = P_{\texttt{Fix}} \,\|\, P_{\texttt{Editable},k}, \tag{16}$$

where $P_{\texttt{Fix}}$ contains static information about the task, input format, and solver API that remains unchanged across iterations, and $P_{\texttt{Editable},k}$ is the dynamic portion, refined at each step using semantic feedback. The update rule mimics gradient-based optimization but operates in the space of text. The verbal feedback from Step 3 serves as the "text-gradient," which is appended to the editable portion of the prompt:

$$P_{\texttt{Editable},k+1} \leftarrow P_{\texttt{Editable},k} + \text{Feedback}_{\text{verbal}}^k, \tag{17}$$

and construct the new prompt as:

$$\texttt{Prompt}_{k+1} = P_{\texttt{Fix}} \,\|\, P_{\texttt{Editable},k+1}. \tag{18}$$

This updated prompt, now containing a historical account of what has been tried and a clear directive for what to do next, is fed to the LLM in the next iteration (Step 1). This iterative refinement loop, summarized in Algorithm 1, continues until a satisfactory trade-off between utility and preference satisfaction is achieved or a set number of episodes is completed.

Remark 4.4. *Our framework treats the LLM as an adaptive reward generator, using trajectory-level comparisons to produce verbal feedback that acts as a soft "gradient" in prompt*

space. This enables iterative refinement of shaping rewards toward better utility-preference trade-offs, without requiring to specify trade-off weights in (Hayes et al. 2022).

Remark 4.5 (Multi-Run Pareto Exploration)**.** *While each execution of* VORTEX *converges to a single Pareto-optimal point (Theorem 5.2 in Section 5), stakeholders may wish to explore different trade-offs between task utility and preference satisfaction. Our framework naturally supports this through multiple runs, where each iteration is conditioned on feedback from the previous result. If stakeholders find the current balance unsatisfactory-for instance, preferring higher coverage despite reduced efficiency-they can provide directional feedback that guides the next run toward a different region of the Pareto frontier. This iterative refinement process allows practical navigation of trade-offs without requiring explicit weight specification.*

Algorithm 1: VORTEX: Verbal-guided Optimization with Reward Tuning via Experiential Trajectory Exploration

Require: Initial $\texttt{Prompt}_0$, task domain, base reward R_{base}, budget B, number of episodes K;

Initialize $\tau^{-1} \leftarrow$ random baseline (or empty trajectory)
for $k = 0$ to $K - 1$ **do**
  (LLM) Generate shaping reward R_h^k following (11);
  (Solver) Solve policy π^k according to (10);
  (Execute) Deploy π^k in $\mathcal{E}$ to collect trajectory τ^k;
  (Compute) Evaluate task utility U_k in (12) and feature distribution $D_k(z)$ in (13);
  if $k > 0$ **then**
    (Compare) Compute difference: δ_U, δ_Δ in (14) ;
    (Feedback) Generate verbal reflection by (15);
    (Text-Gradient) Update prompt based on (18);
  end if
end for
Return: Final $\texttt{Prompt}_K$, policy π^K;

5 Theoretical Guarantee

We analyze the convergence properties of our iterative framework VORTEX. The key challenge is that we are optimizing in the space of reward shaping functions to explore the Pareto frontier. We first make some necessary assumptions, and then present the main result.

5.1 Convergence Analysis

Assumption 5.1. *We make the following assumptions:*

1. ***(A1) Solver Optimality:*** *For given shaping reward R_h, the external solver returns globally optimal policy $\pi^*(R_h)$.*
2. ***(A2) Preference Convexity:*** *The divergence term $C(\pi)$ is convex with respect to the feature distribution.*
3. ***(A3) Text-Gradient Quality:*** *The verbal reinforcement provides an unbiased (or bounded-bias) stochastic estimate of gradient $\nabla_{R_h}[J_\lambda(R_h)]$ with bounded variance.*

Assumption (A1) abstracts away the complexity of the underlying optimization problem by treating the solver as a reliable oracle, realistic in many applications. It allows us to focus the analysis on the behavior of the LLM-driven reward shaping mechanism rather than solver errors. Assumption (A2) ensures the existence of gradients and well-behaved optimization landscape, which can be easily satisfied when f-divergence is KL-divergence or total variation. Assumption (A3) is the most critical, requiring that LLM feedback provides meaningful directional information with vanishing bias, which is supported by the structured nature of trajectory comparisons.

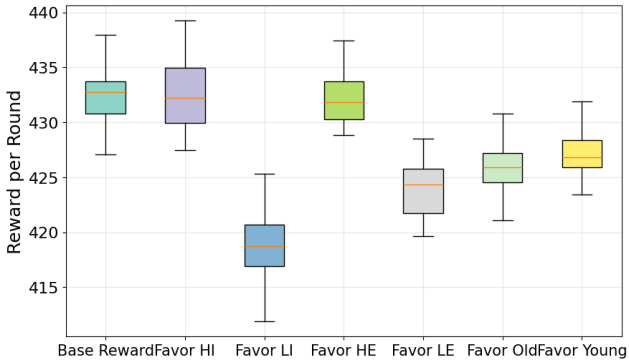


Figure 4: Reward comparison.

5.2 Main Result

Theorem 5.2 (Convergence to Pareto Optimal Point)**.** *If the LLM's preference encoding corresponds to some implicit scalarization weight λ in* (7)*, under Assumptions (A1)–(A3), the proposed iterative* VORTEX *converges almost surely to a stationary point R_h^* such that the resulting policy $\pi^*(R_h^*)$ achieves a Pareto optimal trade-off:*

$$\Big(U(R_h^*) = \mathbb{E}_{\pi^*(R_h^*)}[R_{base}],\ C(R_h^*) = -C(\pi^*(R_h^*))\Big) \in \mathcal{P}.$$

6 Experiments

To evaluate VORTEX, we simulate a constrained public health intervention scenario inspired by the ARMMAN maternal health setting (Mate et al. 2022; Behari et al. 2024) and a conservation setting (Qian et al. 2016). For fair comparison with SOTA baseline DLM in (Behari et al. 2024), we use Gemini-2.5-Pro as the LLM in our experiments. We present the main results for the public health domain in this section, and detailed setting descriptions and more results can be found in Appendix C in the supplementary materials.

Environment abstract. We simulate a population of 800 mothers, evenly partitioned into 8 demographic types based on three binary features: **Income:** Low / High; **Education:** Low / High; **Age:** Young / Old. At each round, the planner is allowed to intervene with up to $B = 400$ mothers. **Preference requirement:** The human decision-maker specifies a soft equity preference such as: *"Slightly prefer mothers with specific features, such as low income, low education, young age."* In this simulation, we consider 6 different preferences as favor high/low income(HI/LI), high/low education (HE/LE), Old, and Young.

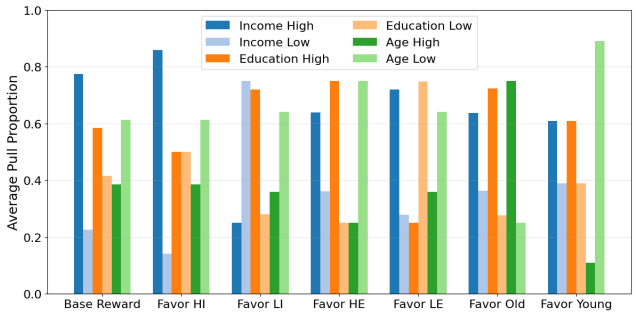
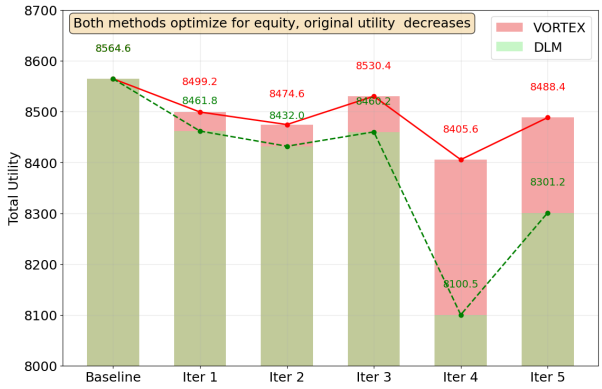


Figure 5: Coverage ratio comparison.

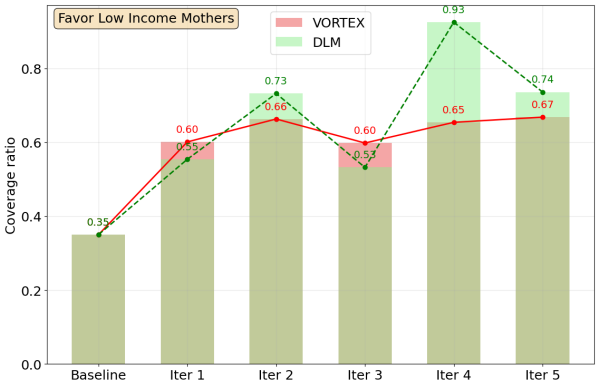
6.1 Results

Effectiveness of Reward Shaping. The effectiveness of the proposed reward shaping technique is evaluated by visualizing the trade-off between task utility and human preference satisfaction.

As shown in Figure 4, the "Base Reward" policy, optimized solely for utility, achieves the highest performance with a median reward of approximately 432. As soon as any human preference is introduced via reward shaping, total utility declines. This result highlights the inherent cost of alignment, demonstrating that satisfying qualitative preferences requires a trade-off with the unconstrained performance metric.



(a) Total utility (favor LI)



(b) Coverage ratio (favor LI)

Figure 6: Comparison with DLM for ARMMAN.

Figure 5 illustrates the success of reward shaping in enforcing preferences. While the "Base Reward" policy is clearly biased (e.g., allocating resources 0.78 to high-income vs. 0.22 to low-income individuals), this is effectively corrected when a preference is applied. For instance, under the "Favor LI" condition, the pull proportion for LI individuals substantially increases to 0.75. This successful shift towards the targeted group is observed across all tested preference conditions.

Baseline comparison. Figure 6 presents a comparison between our proposed VORTEX, against the SOTA baseline, DLM. The comparison is conducted over five iterative rounds, evaluating both total utility and the coverage ratio for one specified human preferences ("Favor LI"). More experiments are relegated to Appendix C.

Figures 6a compares the total utility. VORTEX maintains a high and stable utility after a controlled initial drop, while the DLM baseline is highly volatile and suffers a dramatic drop in utility during its run. This highlights VORTEX's ability to incorporate human preferences without the excessive performance cost and instability exhibited by DLM. Figures 6b shows the coverage ratio for the targeted "low income" group, highlighting the different trade-offs made by each method. VORTEX achieves a steady and stable increase in coverage throughout the iterations. In contrast, DLM's coverage is highly unstable, with erratic fluctuations and a high peak that corresponds to its utility collapse. VORTEX demonstrates a more balanced and reliable performance, substantially improving coverage while preserving high utility.

Pareto front navigation. Figure 7 visualizes how VORTEX navigates the Pareto front, illustrating the trade-off between Task Utility and Preference Satisfaction. It shows two distinct trajectories (T1 and T2) that start from a shared, high-utility baseline. As the iterations progress, both trajectories sacrifice utility to gain preference satisfaction, each exploring a different region of the solution space. Both runs demonstrate clear convergence, settling on different final solutions: T1 favors a higher utility, while T2 prioritizes preference satisfaction. This highlights the framework's ability to converge to stable, well-balanced solutions at different points on the Pareto frontier, catering to varying stakeholder priorities.

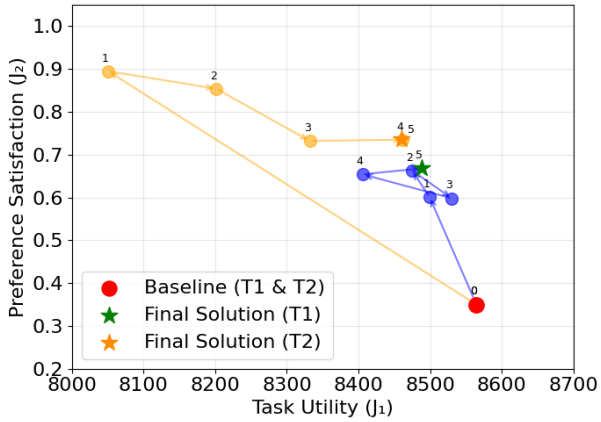


Figure 7: Pareto navigation.

Acknowledgements

This material is based upon work supported by the AI Research Institutes Program funded by the National Science Foundation under the AI Institute for Societal Decision Making (NSF AI-SDM), Award No. 2229881.

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning*.

Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M.; and Robinson, D. G. 2020. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 252–260.

Altman, E. 1999. *Constrained Markov Decision Processes*. Chapman and Hall/CRC.

Bambra, C.; Riordan, R.; Ford, J.; and Matthews, F. 2020. The COVID-19 pandemic and health inequalities. *J Epidemiol Community Health*, 74(11): 964–968.

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. *arXiv preprint arXiv:2402.14807*.

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Conference on Fairness, Accountability, and Transparency*.

Chouldechova, A.; and Roth, A. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Christiano, P.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

Dilkina, B.; Houtman, R.; Gomes, C. P.; Montgomery, C. A.; McKelvey, K. S.; Kendall, K.; Graves, T. A.; Bernstein, R.; and Schwartz, M. K. 2017. Trade-offs and efficiencies in optimal budget-constrained multispecies corridor networks. *Conservation Biology*, 31(1): 192–202.

Fiedrich, F.; Gehbauer, F.; and Rickers, U. 2000. Optimized resource allocation for emergency response after earthquake disasters. *Safety science*, 35(1-3): 41–57.

Hayes, C. F.; Rădulescu, R.; Bargiacchi, E.; Källström, J.; Macfarlane, M.; Reymond, M.; Verstraeten, T.; Zintgraf, L. M.; Dazeley, R.; Heintz, F.; et al. 2022. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36(1): 26.

Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2017. Fairness in reinforcement learning. In *International Conference on Machine Learning*.

Kim, C. W.; Moondra, J.; Verma, S.; Pollack, M.; Kong, L.; Tambe, M.; and Gupta, S. 2025a. Navigating the Social Welfare Frontier: Portfolios for Multi-objective Reinforcement Learning. *arXiv preprint arXiv:2502.09724*.

Kim, C. W.; Verma, S.; Tec, M.; and Tambe, M. 2025b. Preference Robustness for DPO with Applications to Public Health. *arXiv preprint arXiv:2509.02709*.

Kwon, M.; Xie, S. M.; Bullard, K.; and Sadigh, D. 2023. Reward design with language models. *arXiv preprint arXiv:2303.00001*.

Li, H.; Yang, X.; Wang, Z.; Zhu, X.; Zhou, J.; Qiao, Y.; Wang, X.; Li, H.; Lu, L.; and Dai, J. 2024. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16426–16435.

Ma, Y. J.; Liang, W.; Wang, G.; Huang, D.-A.; Bastani, O.; Jayaraman, D.; Zhu, Y.; Fan, L.; and Anandkumar, A. 2023. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv:2310.12931*.

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.

Miettinen, K. 2012. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media.

Ng, A. Y.; Harada, D.; and Russell, S. J. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*.

Ng, A. Y.; and Russell, S. J. 2000. Algorithms for Inverse Reinforcement Learning. In *Proceedings of the 17th International Conference on Machine Learning*.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Pressey, R. L.; Cabeza, M.; Watts, M. E.; Cowling, R. M.; and Wilson, K. A. 2007. Conservation planning in a changing world. *Trends in ecology & evolution*, 22(11): 583–592.

Qian, Y.; Zhang, C.; Krishnamachari, B.; and Tambe, M. 2016. Restless poachers: Handling exploration-exploitation tradeoffs in security domains. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 123–131.

Roijers, D. M.; Vamplew, P.; Whiteson, S.; and Dazeley, R. 2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*.

Shi, Z. R.; Wang, C.; and Fang, F. 2020. Artificial intelligence for social good: A survey.

Venkataraman, S.; et al. 2021. Conservative fairness approaches for policy learning. In *Advances in Neural Information Processing Systems*.

Verma, S.; Boehmer, N.; Kong, L.; and Tambe, M. 2025. Balancing Act: Prioritization Strategies for LLM-Designed Restless Bandit Rewards. *GameSec, arXiv preprint arXiv:2408.12112*.

Yu, W.; Gileadi, N.; Fu, C.; et al. 2023. Language to Rewards for Robotic Skill Synthesis. In *Conference on Robot Learning*.

Yu, Y.; Yao, Z.; Li, H.; Deng, Z.; Jiang, Y.; Cao, Y.; Chen, Z.; Suchow, J.; Cui, Z.; Liu, R.; et al. 2024. Fincon: A

synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37: 137010–137045.