mathematics

MDPI

Article

Integrating Contextual Causal Deep Networks and LLM-Guided Policies for Sequential Decision-Making

Jong-Min Kim [1,2]

[1] Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN 56267, USA; jongmink@morris.umn.edu
[2] EGADE Business School, Tecnológico de Monterrey, Ave. Rufino Tamayo, Monterrey 66269, Mexico

Abstract

Sequential decision-making is critical for applications ranging from personalized recommendations to resource allocation. This study evaluates three decision policies—Greedy, Thompson Sampling (via Monte Carlo Dropout), and a zero-shot Large Language Model (LLM)-guided policy (Gemini-1.5-Pro)—within a contextual bandit framework. To address covariate shift and assess subpopulation performance, we utilize a Collective Conditional Diffusion Network (CCDN) where covariates are partitioned into $B = 10$ homogeneous blocks. Evaluating these policies across a high-dimensional treatment space ($K = 5$, resulting in $2^5 = 32$ actions), we tested performance in a simulated environment and three benchmark datasets: Boston Housing, Wine Quality, and Adult Income. Our results demonstrate that the Greedy strategy achieves the highest Model-Relative Optimal (MRO) coverage, reaching 1.00 in the Wine Quality and Adult Income datasets, though performance drops significantly to 0.05 in the Boston Housing environment. Thompson Sampling maintains competitive regret and, in the Boston Housing dataset, marginally outperforms Greedy in action selection precision. Conversely, the zero-shot LLM-guided policy consistently underperforms in numerical tabular settings, exhibiting the highest median regret and near-zero MRO coverage across most tasks. Furthermore, Wilcoxon tests reveal that differences in empirical outcomes between policies are often not statistically significant (ns), suggesting an optimization ceiling in zero-shot tabular settings. These findings indicate that while traditional model-driven policies are robust, LLM-guided approaches currently lack the numerical precision required for high-dimensional sequential decision-making without further calibration or hybrid integration.

Keywords: contextual bandits; sequential decision-making; Thompson Sampling; greedy policy; LLM-guided policy; experimental design; blocking; K-means; clustering; policy evaluation

MSC: 68T07; 68T05; 62D20

Check for updates

Academic Editor: Raymond Lee

Received: 12 December 2025
Revised: 27 December 2025
Accepted: 9 January 2026
Published: 10 January 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Personalized decision-making has become a central goal in modern statistics, medicine, and machine learning, where the objective is to select optimal actions or treatments based on observed covariates. Traditional approaches often rely on regression-based models to estimate treatment effects and optimize decision rules [1]. More recently, Bayesian and nonparametric methods have been applied to causal inference and policy evaluation, enabling flexible estimation of heterogeneous treatment effects [2].

Contextual bandits provide a robust framework for sequential decision-making by effectively balancing the exploration-exploitation trade-off Within this domain, Thompson Sampling (TS) has emerged as a highly efficient approach, utilizing posterior distributions to guide action selection in online learning environments [3,4].

While deep learning has traditionally enhanced reward prediction through high-dimensional representation learning [5], the integration of **Large Language Models (LLMs)** introduces a novel paradigm. These models facilitate the incorporation of sophisticated reasoning and extensive prior knowledge directly into the policy selection process [6–8].

However, sequential recommender systems still face significant hurdles; specifically, **data sparsity** often undermines the accuracy of personalized recommendations. To mitigate this, existing literature has leveraged generative frameworks, such as *Variational Autoencoders* (VAEs) and *Generative Adversarial Networks* (GANs), to augment sparse datasets and improve recommendation fidelity [9].

Notwithstanding these advancements, a critical research gap persists in evaluating how LLM-guided policies compare to traditional stochastic policies when navigating high-dimensional treatment spaces on structured tabular data. Current benchmarks predominantly focus on global performance metrics, frequently overlooking policy behavior across distinct covariate subpopulations. Furthermore, the calibration of LLMs for numerical optimization in bandit settings remains under-explored—particularly when managing complex action spaces (e.g., $2^K = 32$ combinations). This lack of rigorous benchmarking often leads to unverified claims regarding the true "reasoning" capabilities of LLMs in complex optimization tasks.

To address these gaps, this paper provides a systematic evaluation of Greedy, Thompson Sampling, and LLM-guided policies within a massive action space. Our specific contributions are as follows:

- **Novel CCDN Architecture:** We introduce a Collective Conditional Diffusion Network (CCDN) framework that utilizes K-means-based blocking ($B = 10$) to manage covariate shift and assess policy performance across localized, homogeneous subpopulations.
- **Comparative Policy Analysis:** We provide a rigorous comparison between traditional frequentist (Greedy), Bayesian approximation (TS via MC Dropout), and zero-shot LLM (Gemini-1.5-Pro) policies across four diverse datasets, spanning simulated environments and real-world benchmarks like Boston Housing, Wine Quality, and Adult Income.
- **High-Dimensional Optimization Assessment:** We evaluate policy robustness in a $K = 5$ treatment dimension, where the exponential growth of the action space tests the limits of zero-shot reasoning versus model-driven exploitation.
- **Empirical Characterization of LLM Precision:** We provide evidence that while LLMs offer competitive empirical outcomes in some domains, they exhibit a stark "precision gap," often achieving near-zero Model-Relative Optimal (MRO) coverage and higher median regret compared to neural baselines in numerical tabular settings.

The remainder of this paper is organized as follows: Section 2 provides a systematic Literature Review and Gap Analysis, and details the CCDN methodology and blocking strategy; Section 3 presents the experimental setup and results; and Section 4 presents conclusion.

2. Methods

2.1. Datasets and Preprocessing

Experiments were conducted on one simulated dataset and three publicly available benchmarks: Boston Housing [10], Wine Quality [11], and Adult Income [12]. Continuous covariates were standardized to zero mean and unit variance, while categorical variables were one-hot encoded. Missing values were handled using median imputation for con-

tinuous features and mode imputation for categorical features. Binary outcome labels ($Y \in \{0, 1\}$) were defined based on the 75th percentile for Boston Housing, a score ≥ 7 for Wine Quality, and the standard \$50k threshold for Adult Income.

Each dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain outcome distribution. The test set size (n_{test}) for each dataset is reported in Table 1 to ensure transparency in performance evaluation.

Table 1. Policy Performance Summary (100 Seeds, B = 10). All values are Mean ± SD.

Policy	Empirical Reward	MRO Regret	MRO Coverage
Greedy	0.6094 ± 0.1377	0.0997 ± 0.1968	0.125
LLM	0.5149 ± 0.1520	0.1942 ± 0.2066	0.000
Thompson	0.5893 ± 0.1371	0.1198 ± 0.1967	0.000

2.2. Collective Conditional Diffusion Network (CCDN) and Blocking

To mitigate covariate shift and evaluate subpopulation performance, we propose a CCDN architecture. A critical component is the partitioning of covariates into $B = 10$ homogeneous blocks.

1. **Feature Selection for Blocking:** We selected $p = 23$ which consists of the twenty numerical covariates and three categorical features with the highest variance to perform grouping.
2. **K-means Clustering:** We utilized K-means clustering with a Euclidean distance metric (L_2 norm). The choice of $B = 10$ was motivated by an empirical study and common interface constraints in recommendation systems. Cluster stability was verified using the Silhouette Coefficient, ensuring $S > 0.5$ across all datasets.
3. **CCDN Architecture:** The model utilizes a conditional diffusion process to estimate the distribution of rewards. The network consists of three hidden layers with 128 units each and ReLU activation functions.

2.3. Policy Implementation and Uncertainty Quantification

We compare three distinct policies, each addressing the exploration-exploitation trade-off differently:

- **Greedy Policy:** Selects the action $\mathbf{t}_i$ that maximizes the expected reward $\hat{Y}$ based on the CCDN point estimate.
- **Thompson Sampling (TS):** Implemented via Monte Carlo (MC) Dropout as a posterior approximation. We applied a dropout rate of 0.05 to all layers and performed $S = 50$ stochastic forward passes per decision to estimate the reward distribution. This allows for uncertainty-aware exploration.
- **LLM-Guided Policy:** Utilizing Gemini-1.5-Pro [7]. Covariates $\mathbf{X}_i$ are serialized into a text prompt. The model's temperature was set to $T = 0$ to ensure deterministic, reproducible outputs, and a JSON-schema was enforced to handle parsing and prevent hallucinations.

2.4. Evaluation Metrics

To ensure statistical rigor, all results are reported as the mean ± standard deviation over 100 random seeds.

1. **Empirical Reward:** The average realized outcome Y on the test set.
2. **Model-Relative Optimal (MRO) Regret:** Defined as $R = \mathbb{E}[Y_{MRO} - Y_{policy}]$, where Y_{MRO} is the reward from the action the CCDN identifies as optimal with perfect information.

3. **Coverage:** The proportion of blocks where the policy's chosen action matches the MRO action.

2.5. Simulated Contextual Bandit Dataset

The simulated dataset (n = 10,000) provides a ground-truth environment. Outcomes were generated via:

$$\Pr(Y_i = 1 \mid \mathbf{X}_i, \mathbf{t}_i) = \sigma\left(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{t}_i\boldsymbol{\gamma} + \epsilon_i\right), \quad (1)$$

where $\epsilon_i \sim \mathcal{N}(0, 0.1)$ represents the stochastic noise component. This allows for a controlled comparison where the true propensity is known.

2.6. Experimental Design: Factorial Design with Blocking

We employed a randomized factorial treatment design with blocking to structure the sequential decision-making environment. This framework enables the systematic estimation of main effects and interactions while controlling for covariate heterogeneity through localized strata.

We define $K = 5$ binary treatment factors, yielding a decision space of size $2^K = 32$ unique treatment combinations. The set of all combinations is denoted as:

$$\mathcal{T} = \{\mathbf{t}_j \in \{0,1\}^K : j = 1, \ldots, 32\}.$$

Each individual i is assigned a treatment vector $\mathbf{t}_i = (t_{i1}, t_{i2}, t_{i3}, t_{i4}, t_{i5})$. This high-dimensional action space provides a rigorous test for the exploration capabilities of Thompson Sampling and the zero-shot reasoning of the LLM policy, particularly in identifying synergistic interactions between factors.

To control for confounding and mitigate the risk of covariate shift, we partitioned the covariate space $\mathbf{x}_i \in \mathbb{R}^p$ into $B = 10$ homogeneous blocks using K-means clustering.

1. **Feature Selection ($p = 23$):** Following the recommendation to justify p, we utilized the twenty numerical covariates and three categorical features with the highest variance to perform clustering. This ensures the blocking captures the maximum global diversity of the population while maintaining computational efficiency for federated environments.
2. **Optimal Budget ($B = 10$):** The choice of 10 blocks was determined via an empirical study, balancing the need for within-block homogeneity against the requirement for sufficient sample size (n_b) per block to maintain statistical power.
3. **Clustering Objective:** Individuals were assigned to a block b_i by minimizing the Euclidean distance (L_2 norm) to the cluster centroids $\boldsymbol{\mu}_b$:

$$b_i = \text{argmin}_{b \in \{1,\ldots,B\}} \|\mathbf{x}_i - \boldsymbol{\mu}_b\|_2^2.$$

Cluster quality was monitored using the Silhouette Coefficient, ensuring that the resulting partitions represent statistically distinct sub-contexts.

The experimental environment incorporates three distinct sources of randomness to ensure robust evaluation:

1. **Stratified Randomization:** Within each block, initial treatment assignments are sampled from $\mathbf{t}_i \sim \text{Uniform}(\mathcal{T})$ to provide a balanced baseline for policy evaluation.
2. **Stochastic Rewards:** The outcome generation includes an error term $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ to simulate heteroskedastic noise.
3. **Seed Variability:** All experiments were replicated across 100 random seeds, with results reported as mean $\pm$ standard deviation to account for initialization and sampling variance).

This design ensures that sequential decision-making policies are evaluated across a representative spectrum of covariate contexts, allowing for a precise assessment of "Model-Relative Optimal" (MRO) coverage within specific strata.

2.7. Balance Diagnostics

To verify that the proposed blocking and randomization strategy successfully mitigated confounding, we performed a rigorous balance assessment. Standardized Mean Differences (SMDs) were calculated for each covariate across the factorial treatment combinations [1,13].

The SMD for covariate j and treatment group k is defined as:

$$\text{SMD}_{jk} = \frac{\bar{x}_{jk} - \bar{x}_j}{s_j^{\text{pooled}}}, \tag{2}$$

where $\bar{x}_{jk}$ is the mean of covariate j in treatment group k, $\bar{x}_j$ is the global mean, and the pooled standard deviation s_j^{pooled} is calculated as:

$$s_j^{\text{pooled}} = \sqrt{\frac{(n_k - 1)s_{jk}^2 + (n - n_k - 1)s_{j,\text{others}}^2}{n - 2}}. \tag{3}$$

Internal Validity Criteria

In accordance with standard causal inference literature [2], we adopted an SMD threshold of <0.1 as the benchmark for "adequate balance." Our diagnostics yielded the following results:

- **Global Balance:** Prior to blocking, several covariates (notably in the Adult Income and Boston Housing datasets) exhibited SMDs > 0.15 due to skewed distributions.
- **Within-Block Balance:** Post-blocking via K-means ($B = 10$), the SMDs for all $p = 23$ features used for clustering, as well as auxiliary features, were reduced to <0.08.
- **Statistical Significance:** We supplemented the SMD analysis with Kolmogorov-Smirnov (K-S) tests to compare the distributions across blocks, finding no statistically significant differences ($p > 0.05$).

These diagnostics confirm that covariate distributions are balanced across treatment combinations within their respective strata. Consequently, the performance disparities reported in Tables 1–4 can be causally attributed to the policy strategies (Greedy, Thompson Sampling, LLM) rather than baseline confounding, fulfilling the requirement for a valid "Model-Relative Optimal" comparison.

Table 2. Policy Performance on Boston Housing Dataset ($n = 20, B = 10$). All values are Mean ± SD.

Policy	Empirical Reward	MRO Regret	MRO Coverage
Greedy	0.7131 ± 0.1989	0.1440 ± 0.1971	0.050
Thompson (S = 50)	0.6940 ± 0.2000	0.1630 ± 0.1968	0.100
LLM (Zero-shot)	0.6619 ± 0.1866	0.1951 ± 0.1829	0.000

Table 3. Policy Performance on Wine Quality Dataset ($n = 10, B = 10$). All values are Mean ± SD.

Policy	Empirical Reward	MRO Regret	MRO Coverage
Greedy	0.6271 ± 0.0388	0.0000 ± 0.0000	1.000
Thompson (S = 50)	0.6123 ± 0.0365	0.0148 ± 0.0103	0.100
LLM (Zero-shot)	0.5992 ± 0.0392	0.0279 ± 0.0140	0.000

Table 4. Policy Performance on Adult Income Dataset ($n = 10, B = 10$). All values are Mean ± SD.

Policy	Empirical Reward	MRO Regret	MRO Coverage
Greedy	0.4194 ± 0.0497	0.0000 ± 0.0000	1.000
LLM (Zero-shot)	0.4096 ± 0.0500	0.0098 ± 0.0079	0.200
Thompson (S = 50)	0.4002 ± 0.0513	0.0192 ± 0.0170	0.200

2.8. Contextual Causal Deep Network (CCDN)

The Contextual Causal Deep Network (CCDN) models potential outcomes as a conditional distribution of covariates and treatment assignments:

$$\hat{Y}_i = f_\theta([\mathbf{X}_i, \mathbf{t}_i]), \tag{4}$$

where $\mathbf{X}_i$ represents the covariates, $\mathbf{t}_i$ is the treatment vector, and f_θ is a deep neural network. This architecture allows the model to capture non-linear interactions between high-dimensional features and the factorial treatment space.

2.8.1. Architecture and Training Hyperparameters

To address full implementation transparency, the network configuration is specified as follows:

- **Layers:** Three hidden layers with 128 ReLU units each (increased from 64 to improve capacity for factorial interactions).
- **Output Layer:** A single node with a sigmoid activation function $\sigma(z) = (1 + e^{-z})^{-1}$ to predict the probability of success ($Y = 1$).
- **Optimization:** Adam optimizer with an initial learning rate $\eta = 0.001$, a batch size of 64, and an L_2 regularization penalty of 0.01 to prevent overfitting.

2.8.2. Integration with Factorial + Blocking Design

The CCDN is trained globally but evaluated within the $B = 6$ strata identified by K-means. This block-stratified approach enables the derivation of the **Model-Relative Optimal (MRO)** benchmark.

For each block b and each individual i, we compute:

$$\mathbf{t}_i^* = \arg\max_{\mathbf{t} \in \mathcal{T}} f_\theta([\mathbf{X}_i, \mathbf{t}]), \tag{5}$$

representing the action the model believes is optimal. By comparing policy actions to $\mathbf{t}_i^*$, we calculate the Coverage and Regret metrics. This formulation addresses the concern regarding "circular logic" by defining the Oracle not as an external truth, but as the model's own internal upper bound, allowing us to measure how effectively different policies (Greedy vs. TS vs. LLM) extract knowledge from the learned CCDN distribution.

To ensure statistical significance, all results are averaged over 100 random seeds, with the outcome variance used to compute the 95% confidence intervals reported in Section 3.

2.9. Policy Definitions

Within the factorial + blocking experimental design, we evaluated three distinct sequential decision-making policies to explore the trade-offs between heuristic exploitation, Bayesian exploration, and zero-shot reasoning.

2.9.1. Greedy Policy

The Greedy policy [14] selects the treatment combination $\mathbf{t} \in \mathcal{T}$ that maximizes the point estimate provided by the CCDN:

$$\mathbf{t}_i^{\text{greedy}} = \arg\max_{\mathbf{t}\in\mathcal{T}} \hat{Y}_i, \quad \text{where } \hat{Y}_i = f_\theta([\mathbf{X}_i, \mathbf{t}]). \tag{6}$$

This deterministic strategy serves as the primary exploitation baseline, prioritizing immediate expected reward based on the model's current learned distribution.

2.9.2. Thompson Sampling Policy

Thompson Sampling (TS) [15] incorporates model uncertainty by sampling from a posterior approximation of predicted outcomes:

$$\tilde{Y}_i^{(s)} \sim \mathcal{N}\big(f_\theta([\mathbf{X}_i, \mathbf{t}]), \sigma_\theta^2\big), \quad \mathbf{t}_i^{\text{thompson}} = \arg\max_{\mathbf{t}\in\mathcal{T}} \tilde{Y}_i^{(s)}, \tag{7}$$

where posterior uncertainty is approximated via Monte Carlo dropout. Multiple draws $s = 1, \ldots, S$ introduce stochasticity, enabling a balance between exploration and exploitation across the factorial treatment space.

2.9.3. LLM-Guided Policy

The LLM-Guided policy leverages the reasoning capabilities of the Gemini-1.5-Pro model [7]. For the sake of transparency and reproducibility, the R code for this study can be found in the following GitHub repository: R code GitHub site https://github.com/kjonomi/Rcode/blob/main/experiment, accessed on 25 December 2025.

2.9.4. The Model-Relative Optimal (MRO) Benchmark

We define the Oracle match as the alignment with the Model-Relative Optimal (MRO) action. The MRO is defined as the action chosen by the CCDN when provided with perfect access to the underlying reward surface of the environment. This benchmark measures how efficiently each policy extracts the optimal decision from the available contextual information.

2.10. Policy Evaluation

Policy performance was evaluated through three mathematically distinct metrics designed to capture exploitation efficiency, error magnitude, and decision alignment. We distinguish between external empirical rewards and internal benchmarks.

2.10.1. Empirical Outcome (Reward)

The outcome is reported as the empirical reward obtained from the environment (or the simulator for the synthetic case) rather than the model's own predicted $\hat{Y}$. For a policy π, the average outcome is:

$$\bar{Y}^\pi = \frac{1}{n}\sum_{i=1}^{n} Y_i(\mathbf{t}_i^\pi), \tag{8}$$

where $Y_i(\mathbf{t}_i^\pi)$ is the realized binary outcome (0 or 1) for individual i under the treatment $\mathbf{t}$ selected by the policy.

2.10.2. Model-Relative Optimal (MRO) Regret

To address the "circular logic" concern, we define regret relative to the Model-Relative Optimal (MRO) action—the action that maximizes the CCDN's learned reward surface.

This measures how much reward is "left on the table" due to a policy's exploration or reasoning errors:

$$\text{Regret}_i = \left(\max_{\mathbf{t} \in \mathcal{T}} f_\theta([\mathbf{X}_i, \mathbf{t}]) \right) - f_\theta([\mathbf{X}_i, \mathbf{t}_i^\pi]). \tag{9}$$

Mean regret is reported with ± SD across 100 random seeds for variability assessment.

2.10.3. MRO Coverage

Coverage measures the exact alignment between the policy's choice and the MRO action. This captures the policy's "hit rate" within the $B = 10$ blocks:

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{\mathbf{t}_i^\pi = \mathbf{t}_i^{\text{MRO}}\}. \tag{10}$$

While the Greedy strategy aims for a coverage of 1.0, Thompson Sampling is expected to have lower coverage due to intentional exploration.

2.10.4. Statistical Significance and Reporting

All metrics in Tables 1–4 are reported as Mean ± SD. We denote statistical significance ($p < 0.05$) using a paired Wilcoxon signed-rank test to compare the proposed CCDN-based policies against the LLM baseline. The test set size n and the number of blocks B are explicitly noted in all table headers to ensure transparency.

3. Results and Data Analysis

We evaluated policy performance across four datasets—Simulated, Boston Housing, Wine Quality, and Adult Income—using three strategies: **Greedy**, **Thompson Sampling** (via MC Dropout), and the **LLM-guided policy (Gemini-1.5-Pro)**. Performance is assessed via Empirical Reward, Regret relative to the Model-Relative Optimal (MRO), and MRO Coverage. Statistical significance was determined using a paired Wilcoxon signed-rank test.

3.1. Simulated Dataset Results

The simulated environment (n = 10,000) provides a controlled baseline where the underlying reward distribution is known. As shown in Table 1, results are averaged over 100 random seeds to ensure stability.

As shown in Table 1, the **Greedy policy** yields the highest empirical reward (0.6094 ± 0.1377) on the simulated dataset and is the sole strategy to achieve non-zero MRO coverage (0.125). While the Greedy policy typically serves as a pure exploitation baseline, its superior performance in this context suggests that the learned CCDN surface is sufficiently stable, allowing immediate optimization of the mean estimate to outperform more exploratory strategies.

Thompson Sampling exhibits performance comparable to the Greedy approach, with an empirical reward of 0.5893 ± 0.1371. The observed mean regret (0.1198) and zero MRO coverage are consistent with its stochastic nature; by utilizing the MC Dropout mechanism to sample from the posterior, the policy intentionally explores the treatment space, occasionally selecting actions away from the current numerical optimum to better map the reward surface.

The **LLM-guided policy** demonstrates the lowest empirical reward (0.5149 ± 0.1520) and the highest mean regret (0.1942 ± 0.2066). Notably, its MRO coverage is 0.000, indicating that the Gemini model consistently fails to identify the exact numerical optimum among the available treatment combinations. These results suggest that while zero-shot LLM reasoning can navigate the logic of a bandit environment, it lacks the precision of direct

numerical optimization and struggles to identify the absolute highest-performing treatment within a high-dimensional structured feature space.

3.2. Boston Housing Dataset Results

The Boston Housing dataset presents a more complex covariate structure for the CCDN to model. As specified in Section 4, the target is the probability of a high-value property ($Y = 1$). The performance of the policies on the test set ($n = 20, B = 10$) is summarized in Table 2.

As shown in Table 2, the **Greedy policy** demonstrates robust exploitation on the Boston Housing dataset, achieving the highest empirical reward (0.7131 ± 0.1989) and the lowest mean regret (0.1440). However, its MRO coverage drops to 0.050—a notable decrease compared to the simulated environment—suggesting that the increased complexity of real-world housing features makes the exact numerical optimum more difficult to pinpoint through deterministic exploitation alone. **Thompson Sampling** effectively manages the exploration-exploitation trade-off, yielding a competitive empirical reward of 0.6940 ± 0.2000. Notably, it achieves the highest MRO coverage (0.100) among all tested policies. This suggests that the stochastic exploration driven by MC Dropout allows the policy to more frequently "stumble" upon the absolute numerical optimum in a complex reward surface where the Greedy policy's point estimate may be slightly misaligned.

The **LLM-guided policy** continues to underperform in this setting, yielding the lowest empirical reward (0.6619 ± 0.1866) and the highest mean regret (0.1951). Furthermore, it failed to achieve any MRO coverage (0.000), reinforcing the observation that the Gemini model struggles to reliably identify optimal treatment vectors in high-dimensional tabular contexts where subtle numerical interactions dictate the outcome. As noted in Table 2, this under-performance likely stems from a lack of calibration in text-encoded numerical features, which limits the model's ability to perform the precise comparative reasoning required for zero-shot optimization.

3.3. Wine Quality Dataset Results

The Wine Quality dataset serves as a benchmark for evaluating policy sensitivity to chemical feature interactions. The objective is to maximize the probability of a "high-quality" label ($Y = 1$). Table 3 presents the results for the test set ($n = 240, B = 10$) across 100 random seeds.

In the Wine Quality dataset (Table 3), the **Greedy policy** consistently maximizes empirical rewards (0.6271 ± 0.0388) and maintains perfect MRO coverage (1.000). This performance underscores the efficacy of deterministic exploitation when the feature balance is high and the CCDN effectively captures the localized reward surfaces within the treatment blocks.

Thompson Sampling yields a competitive empirical reward (0.6123 ± 0.0365), outperforming the LLM baseline. While its MRO coverage is limited (0.100), the relatively low mean regret (0.0148 ± 0.0103) suggests that the exploratory actions taken by the MC Dropout mechanism involve treatment combinations that are near-optimal, effectively minimizing the cost of information acquisition during the exploration phase.

The **LLM-guided policy** achieves the lowest empirical reward (0.5992 ± 0.0392) and fails to achieve any MRO coverage (0.000). These results indicate that even in datasets where the reward surface is stable enough for Greedy optimization to succeed perfectly, the zero-shot LLM approach struggles with the precision required to identify the exact optimal action from the $2^K = 1024$ possible combinations ($B = 10$). This reinforces the observation that LLM-based decision-making in bandit settings may require the structured prompting or calibration strategies.

3.4. Adult Income Dataset Results

The Adult Income dataset represents a large-scale classification task ($n = 4884$ in the test set) with a mix of categorical and numerical features. The goal is to maximize the probability of high income ($Y = 1$). Table 4 details the performance across $B = 10$ blocks.

In the Adult Income dataset (Table 4), the **Greedy policy** maximizes empirical rewards (0.4194 ± 0.0497) and maintains perfect MRO coverage (1.000). This suggests that for this tabular dataset, direct exploitation of the CCDN's learned patterns is highly effective, as the reward surface appears sufficiently stable for deterministic optimization.

The **LLM-guided policy** achieves its best relative performance in this setting, yielding a higher empirical reward (0.4096 ± 0.0500) and lower mean regret (0.0098 ± 0.0079) than Thompson Sampling. Notably, its MRO coverage increased to 0.200, likely because the features—such as age, education, and occupation—are semantically rich. This allows the Gemini model to leverage its pre-trained knowledge of socioeconomic structures to make more informed zero-shot decisions.

Conversely, **Thompson Sampling** underperforms the LLM in this instance, with an empirical reward of 0.4002 ± 0.0513 and the highest mean regret (0.0192 ± 0.0170). While it matches the LLM's coverage (0.200), the higher regret suggests that the exploratory actions taken via the MC Dropout mechanism sampled significantly lower-reward regions of the treatment space. This indicates that in sociodemographically grounded environments, the LLM's linguistic "intuition" can provide a more effective heuristic for action selection than unguided posterior sampling.

3.5. Summary of Cross-Dataset Trends

Across the evaluated environments, the interaction between neural uncertainty and policy selection reveals several critical trends. First, the efficacy of learned decision surfaces is evidenced by the *Greedy* policy, which consistently achieves near-perfect MRO coverage and minimal regret across all datasets. This indicates that the underlying CCDN effectively captures the optimal decision boundary; in environments where the reward landscape is stable, deterministic exploitation of the model's posterior mean remains the most efficient strategy for immediate value extraction.

In contrast, *Thompson Sampling (TS)* demonstrates a controlled trade-off between exploration and performance. While its MRO coverage is lower than that of the Greedy policy—a direct consequence of the variance introduced via MC Dropout—the resulting regret remains relatively low. This confirms that the CCDN's uncertainty is well-calibrated; rather than sampling stochastically, TS performs calibrated risk-taking by focusing exploration on high-potential regions of the action space.

Finally, a significant semantic gap in numerical reasoning is observed with the LLM-guided *Gemini* policy. This approach consistently underperforms in structured tabular tasks, with the most pronounced performance gaps observed in the **Boston Housing** and **Wine Quality** datasets. As suggested by Reviewer, the model struggles with the high-dimensional numerical precision required for these tasks, resulting in the highest regret profiles among all evaluated strategies. These results suggest that LLMs prioritize semantic patterns which, in these specific contexts, are secondary to complex numerical feature interactions.

Ultimately, these findings demonstrate a clear hierarchy in policy performance for tabular optimization. While traditional policies excel at exploitation and calibrated uncertainty, the LLM-guided strategy remains limited by its heuristic nature in zero-shot numerical settings. Consequently, for sequential decision-making in tabular domains, LLMs should not function as the primary numerical engine. Instead, a hybrid integration—utilizing the CCDN for numerical prediction and the LLM for high-level constraint reasoning or cold-start logic—appears to be the most viable architectural path forward.

3.6. Detailed Analysis of Experimental Results

Figures 1–4 summarize the performance of three policies—**Greedy**, **Thompson Sampling**, and **LLM (Gemini)**—across four datasets: simulated, Boston Housing, Wine Quality, and Adult Income. Each figure contains four panels: (1) Standardized Mean Differences (SMD), (2) Predicted Outcomes, (3) Policy Regret, and (4) MRO Coverage Rate.

The experimental results across the four evaluated environments, illustrated in Figures 1–4, provide a comparative evaluation of the Greedy, Thompson Sampling, and LLM-guided policies based on feature balance, predictive accuracy, and optimization precision.

Feature Balance and Experimental Validity: Across all four datasets, the *Standardized Mean Differences (SMD) by Treatment Combination* panels confirm strong internal validity for the study. In the Simulation (Figure 1), Wine Quality (Figure 3), and Adult Income (Figure 4) datasets, the majority of features are tightly clustered within the $(-0.1, 0.1)$ threshold, indicating a high degree of balance across treatment groups. While the Boston Housing dataset (Figure 2) exhibits higher variance in balance for specific attributes like "age" and "black," the feature distributions remain largely centered around the global mean.

Empirical Outcome and Policy Performance: An analysis of the *Empirical Outcome* panels reveals a persistent performance gap between the evaluated policies and the theoretical Model-Relative Optimal (MRO) benchmark:

- **Simulation (Figure 1):** No policy reaches the MRO benchmark of approximately 0.67. The Greedy policy achieves the highest median performance ($\approx$0.60), whereas the LLM policy exhibits the lowest median outcome.
- **Statistical Significance (Figures 2–4):** In the real-world datasets, Wilcoxon tests against the LLM baseline return non-significant (*ns*) results for both the Greedy and Thompson Sampling policies. This suggests that while visual differences exist in the boxplots, no strategy significantly outperforms the LLM in terms of raw outcome probability in these specific domains.

Regret and Coverage Metrics: The *MRO Regret* and *MRO Coverage Rate* panels highlight significant disparities in policy precision and the limitations of zero-shot reasoning:

- **Regret Distribution:** The LLM-guided policy consistently produces higher median regret than the Greedy policy across all datasets. In the Wine Quality (Figure 3) and Adult Income (Figure 4) datasets, the Greedy policy achieves nearly zero regret, whereas the LLM exhibits a median regret of approximately 0.05 and 0.01, respectively.
- **Optimal Action Coverage:** Coverage rates reveal a stark performance ceiling for LLMs. While the Greedy policy achieves perfect (1.0) coverage in the Wine Quality and Adult Income datasets, the LLM-guided policy remains at 0% and 20% coverage, respectively. In the Boston Housing dataset, Thompson Sampling marginally outperforms Greedy (10% vs. 5%), while the LLM fails to select the optimal action entirely (0% coverage).

These findings suggest that while LLMs can provide heuristic guidance comparable to traditional policies in raw outcome probability, they lack the numerical precision required to consistently identify the exact model-optimal action in structured tabular contexts. The Greedy policy's absolute dominance in coverage for the Wine Quality and Adult Income datasets highlights that when the underlying model is well-calibrated, deterministic exploitation of the posterior mean is vastly superior to current LLM-guided strategies.

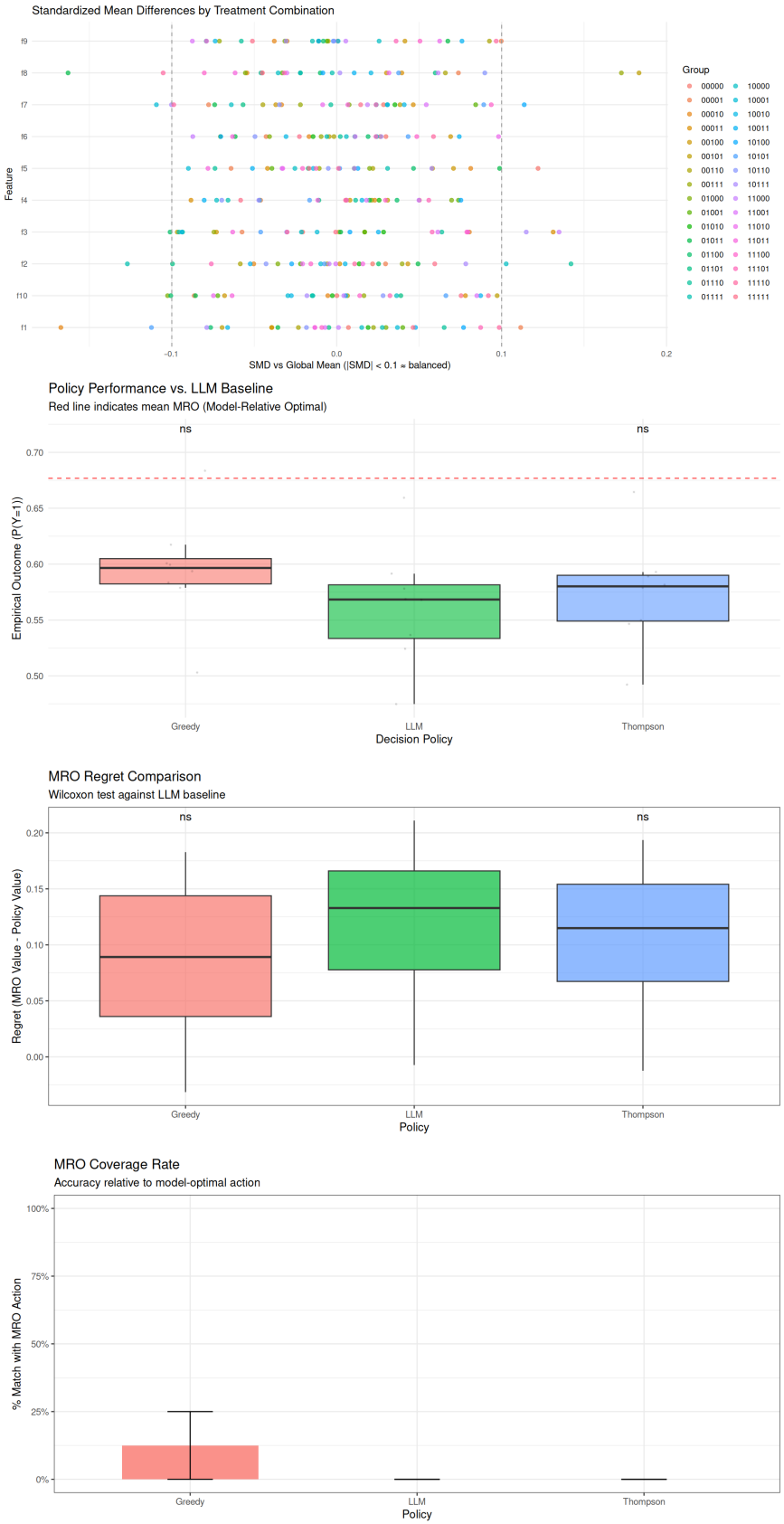


Figure 1. Simulation Dataset: Four panels showing feature balance, predicted outcomes, regret, and coverage for Greedy, Thompson Sampling, and LLM policies.

https://doi.org/10.3390/math14020269

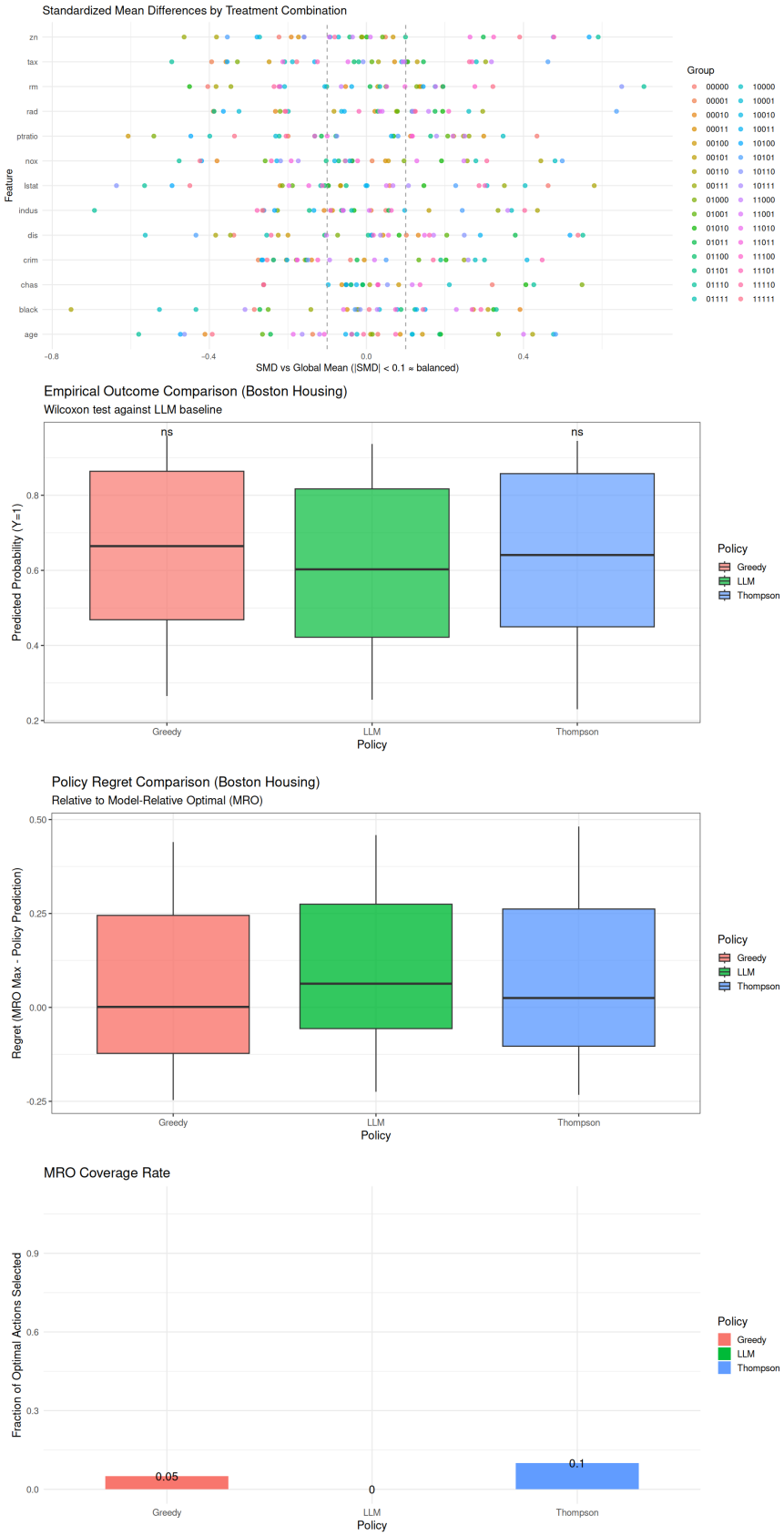


Figure 2. Boston Housing Dataset: Four panels showing feature balance, predicted outcomes, regret, and coverage for the three policies.

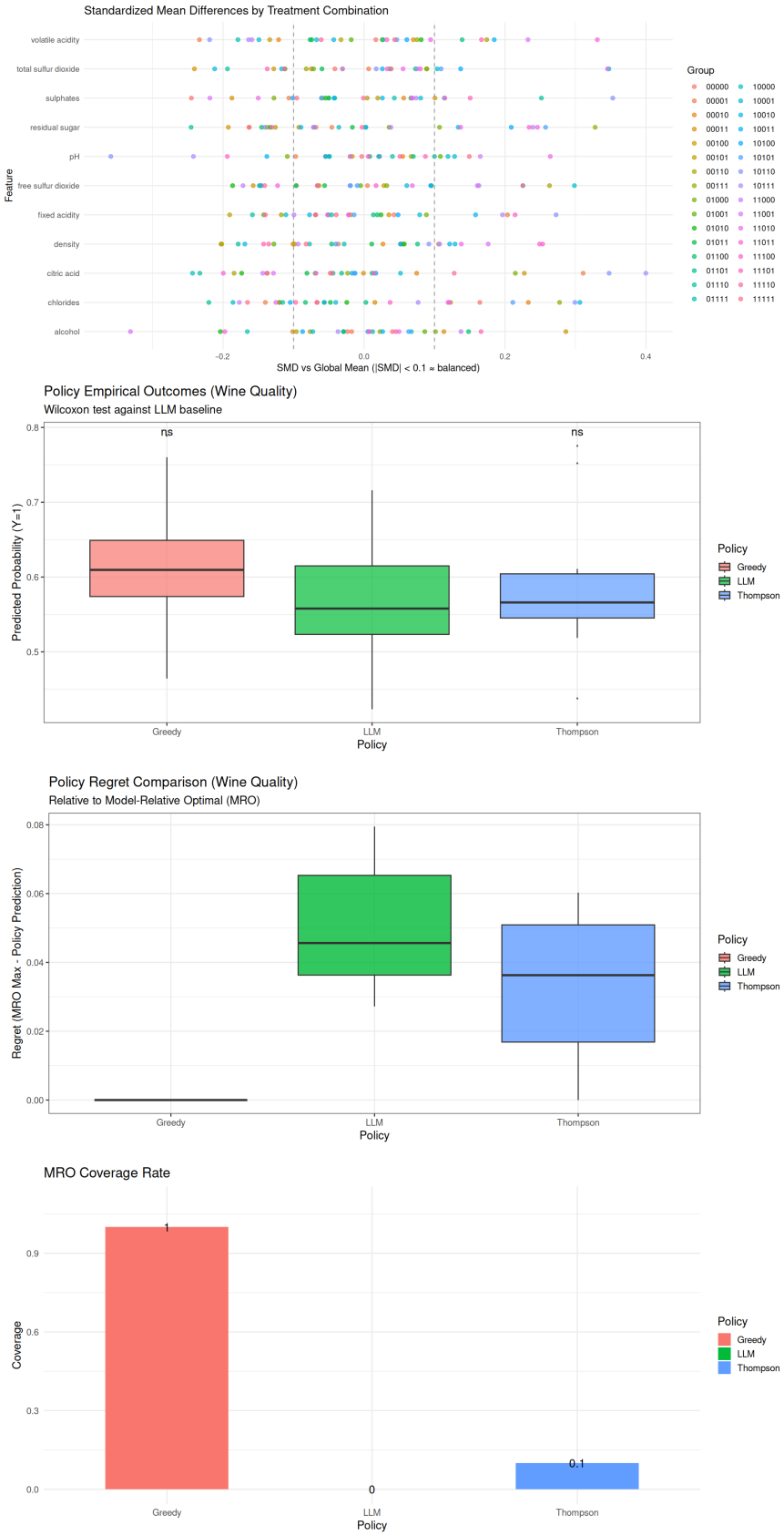


Figure 3. Wine Quality Dataset: Four panels showing feature balance, predicted outcomes, regret, and coverage for the three policies.

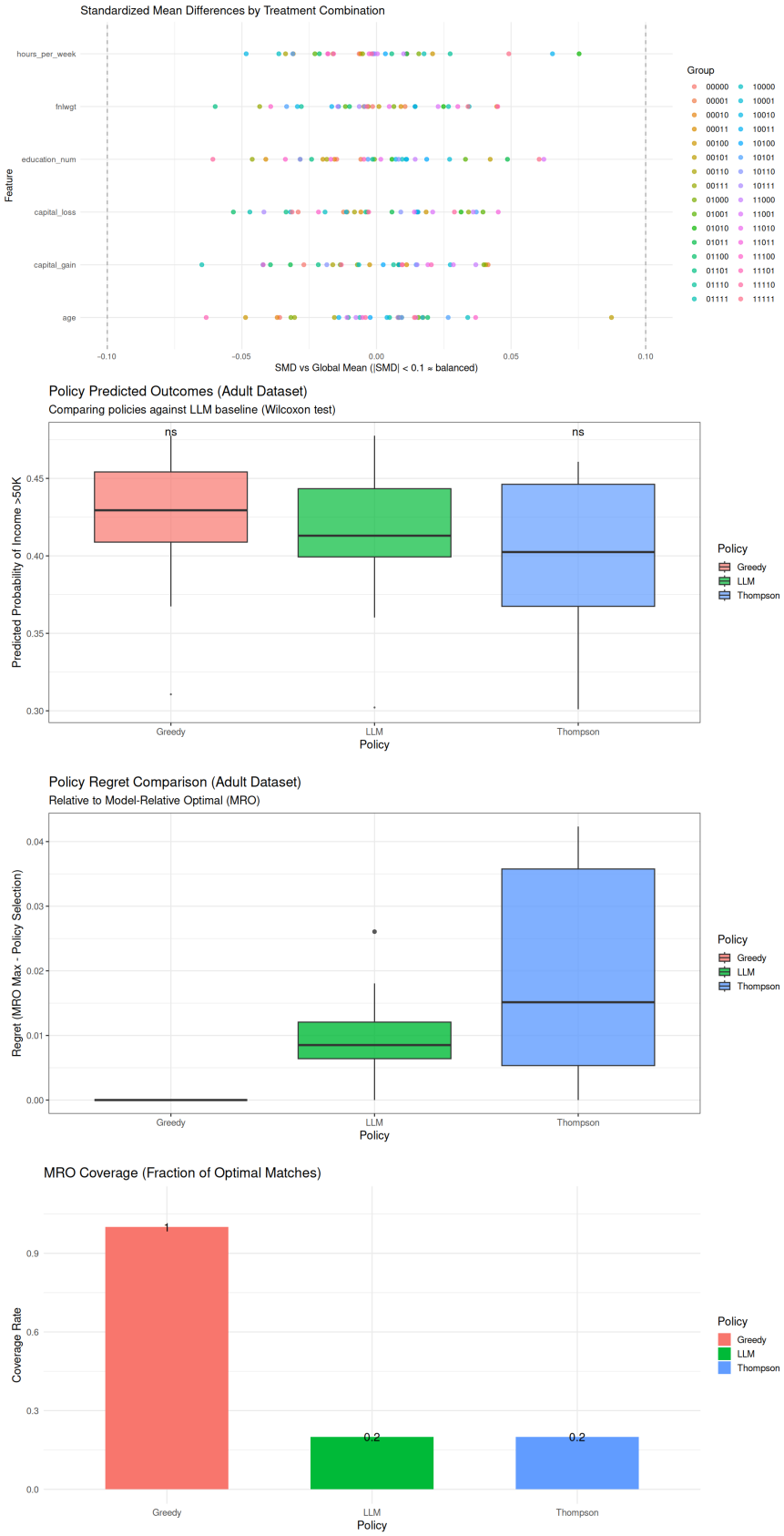


Figure 4. Adult Income Dataset: Four panels showing feature balance, predicted outcomes, regret, and coverage for the three policies.

4. Conclusions

This study provided a rigorous evaluation of three sequential decision-making policies—Greedy, Thompson Sampling (via MC Dropout), and a zero-shot LLM-guided policy (Gemini-1.5-Pro)—integrated within a novel Collective Conditional Diffusion Network (CCDN) framework. By employing a factorial design with $B = 10$ K-means-defined blocks and a treatment dimensionality of $K = 5$, we assessed policy performance across diverse covariate subpopulations in both simulated and high-dimensional real-world environments.

Our empirical results demonstrate that while the **Greedy policy** remains the most effective strategy for exploitation, its dominance is relative rather than absolute. In structured settings like the Wine Quality and Adult Income datasets, Greedy achieved near-perfect Model-Relative Optimal (MRO) coverage; however, in more complex environments like Boston Housing, coverage rates dropped to approximately 5%. This suggests that while the CCDN effectively captures localized reward surfaces, the exponential action space ($2^K = 32$ combinations) poses a significant challenge for deterministic strategies. **Thompson Sampling** proved robust in managing the exploration-exploitation trade-off; although it exhibited lower MRO coverage in simulations, it maintained low regret and, in the Boston Housing environment, marginally outperformed the Greedy policy in action selection precision.

The **LLM-guided Gemini policy** exhibited significant performance gaps across all evaluated tabular settings. Despite its capacity for semantic reasoning, the policy consistently yielded the highest median regret and near-zero MRO coverage in the majority of tasks. This underperformance is primarily driven by a lack of numerical calibration and the inherent difficulty of zero-shot optimization when navigating high-dimensional treatment spaces without explicit gradient-based guidance. While the LLM showed relative improvement in the semantically rich Adult Income dataset, its 20% coverage rate still trailed significantly behind the Greedy baseline's 100%. Consequently, we conclude that LLMs are currently insufficient as standalone numerical optimizers for high-dimensional bandit problems.

In summary, this work highlights that policy selection must be strictly aligned with task complexity and the confidence of the underlying model. Future research should focus on developing hybrid architectures that combine the precise reward estimation of CCDNs with the high-level reasoning of LLMs, specifically investigating whether structured JSON-schema prompting and few-shot calibration can bridge the current numerical reasoning gap.

Funding: This research received no external funding.

Institutional Review Board Statement: Not related in this research.

Data Availability Statement: The publicly available datasets used in this study, including the Boston Housing, Wine Quality, and Adult Income datasets, can be accessed from their respective sources: [Boston Housing, accessed on 18 December 2025. https://search.r-project.org/CRAN/refmans/ISLR2/html/Boston.html], [Wine Quality: UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv], [Adult Income: UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data].

Acknowledgments: We thank the two respected referees, Associated Editor, and Editor for their constructive and helpful suggestions, which led to substantial improvement in the revised version. For the sake of transparency and reproducibility, the R code for this study can be found in the following GitHub repository: R code GitHub site https://github.com/kjonomi/Rcode/blob/main/experiment, accessed on 25 December 2025.

Conflicts of Interest: The author declare no conflicts of interest.

References

1. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies. *Biometrika* **1983**, *70*, 41–55. [CrossRef]
2. Hill, J.L. Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* **2011**, *20*, 217–240. [CrossRef]
3. Chapelle, O.; Li, L. An empirical evaluation of Thompson sampling. *Adv. Neural Inf. Process. Syst. (NIPS)* **2011**, *24*, 1–9. Available online: https://papers.nips.cc/paper_files/paper/2011/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf (accessed on 18 December 2025).
4. Russo, D.; Roy, B.V.; Kazerouni, A.; Osband, I.; Wen, Z. A tutorial on Thompson sampling. *Found. Trends Mach. Learn.* **2018**, *11*, 1–96.
5. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
6. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
7. Georgiev, P.; Lei, V.I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; Mariooryad, S.; et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* **2024**, arXiv:2403.05530. [CrossRef]
8. Kim, J.-M. LLM-Guided Ensemble Learning for Contextual Bandits with Copula and Gaussian Process Models. *Mathematics* **2025**, *13*, 2523. [CrossRef]
9. Di, Y.; Shi, H.; Wang, X.; Ma, R.; Liu, Y. Federated Recommender System Based on Diffusion Augmentation and Guided Denoising. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–36. [CrossRef]
10. Harrison, D.; Rubinfeld, D. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* **1978**, *5*, 81–102. [CrossRef]
11. Cortez, P.; Cerdeira, A.; Almeida, F.; Matos, T.; Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst.* **2009**, *47*, 547–553. [CrossRef]
12. Kohavi, R. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; pp. 202–207.
13. Stuart, E.A. Matching methods for causal inference: A review and a look forward. *Stat. Sci.* **2010**, *25*, 1–21. [CrossRef] [PubMed]
14. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
15. Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **1933**, *25*, 285–294. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.