STRATUM-AWARE LLM REASONING UNDER PER-USER SLOT CONSTRAINTS

Shijin Zhang
Imperial College London
London, United Kingdom
shijin.zhang24@imperial.ac.uk

Tianyu Xia*
Peking University
Beijing, China
2311110185@bjmu.edu.cn

ABSTRACT

In recommendation environments with per-user resource constraints, allocating a limited display slot to an item that a user would have purchased without intervention leads to promotional waste. While traditional uplift modeling estimates the incremental impact of a recommendation, it struggles to operate directly on the unstructured textual data common in modern user histories and item profiles. Furthermore, directly prompting Large Language Models (LLMs) for causal effect estimation often results in unstable numerical reasoning. To address this, we propose **CURA-CLM**, a lightweight end-to-end framework that bridges textual data and strict per-user slot constraints. Our approach first extracts structured causal features from unstructured text using LLMs on a computation-manageable candidate set. We then apply robust statistical estimators to identify the Conditional Average Treatment Effect (CATE). Finally, we introduce a stratum-aware allocation objective that maximizes net revenue by explicitly penalizing resource allocation to non-persuadable user groups. Experiments on synthetic datasets demonstrate that CURA-CLM outperforms direct LLM and standard uplift ranking under per-user capacity constraints.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly used as reasoning components in recommender systems, interpreting user histories, reviews, item descriptions, and interaction narratives for personalized decision-making (Ren et al., 2024; Bao et al., 2023; Shi et al., 2025b; Yue et al., 2025; Wu et al., 2024; Shehmir & Kashef, 2025). Yet practical recommendation is not merely semantic preference matching: candidate items far exceed visible capacity, while user attention and recommendation slots are limited (Jiang et al., 2026). Thus, LLM-assisted recommendation needs to identify which user-item interactions deserve scarce slots, since recommending an item that a user would purchase anyway consumes capacity without generating increments (Li et al., 2023b).

This slot-scarcity setting requires causal rather than purely predictive reasoning. Causal uplift modeling estimates the incremental effect of an intervention (Funk et al., 2005; Künzel et al., 2019; Athey & Wager, 2017; Saito et al., 2019; Zhong et al., 2022), whereas conventional recommendation pipelines optimize post-exposure outcomes, such as click-through rate or purchase probability, without estimating whether recommendation actually causes conversion (Koren et al., 2009; Rendle et al., 2009; He et al., 2017; Cheng et al., 2016; Guo et al., 2018). Existing constrained uplift methods mainly address aggregate promotional expenditure, global budget control, or platform-level cost-benefit ranking (Goldenberg et al., 2020; Sun et al., 2024; Kamran et al., 2024; Goldenberg et al., 2025). These formulations are useful but insufficient for localized interface constraints, where each user has only a few visible slots and one low-yield assignment creates immediate opportunity cost. Therefore, recommendation systems must estimate uplift and allocate estimated effects under hard per-user capacity limits.

This makes recommendation allocation a stratum-aware counterfactual reasoning problem rather than a pure ranking problem. A high predicted conversion probability does not necessarily imply

*Corresponding author.

causal usefulness. From a causal perspective, user-item pairs may belong to latent strata: Persuadables, who purchase only when recommended; Always-Buyers, who purchase regardless; and Never-Buyers, who do not purchase even when recommended (Angrist et al., 1996; Li et al., 2023b). Persuadables generate actural incremental value, Always-Buyers and Never-Buyers waste slots; failing to distinguish them under per-user slot constraints leads to a waste-of-slots problem.

Modern recommendation evidence is increasingly textual: reviews, item descriptions, behavioral histories, and interaction narratives contain signals about preferences, context, value, and recommendation cost. LLMs are well suited for extracting such semantic information, and recent LLM-based recommender systems show strong ability in preference understanding and textual reasoning (Ren et al., 2024; Bao et al., 2023; Shi et al., 2025b; Yue et al., 2025; Wu et al., 2024; Shehmir & Kashef, 2025). However, directly using LLMs for numerical causal estimation or end-to-end counterfactual allocation can be unreliable, especially for causal arithmetic, treatment-effect estimation, or counterfactual comparison (Jin et al., 2023; Liu et al., 2024b; Chen et al., 2023; Jin et al., 2024; Kiciman et al., 2023). Conversely, robust statistical estimators require structured covariates and cannot directly operate on raw text. We therefore use LLMs as semantic-to-structured modules that convert textual evidence into causal covariates for treatment-effect estimators and a slot-constrained optimizer. This follows recent text-to-table causal paradigms but adapts them to localized recommendation scarcity (Liu et al., 2024a; Škrlj et al., 2024; Shi et al., 2025a), allowing LLMs to perform language-based causal abstraction while numerical estimation and allocation remain grounded in statistical modeling.

To bridge these gaps, we propose Causal Uplift-Driven Recommendation Allocation with Causal Large Language Models (**CURA-CLM**), a lightweight end-to-end framework for stratum-aware LLM-assisted recommendation under per-user slot constraints. CURA-CLM combines: (1) LLM-driven structured causal covariate extraction from a computation-manageable candidate set; (2) robust CATE estimation; and (3) a slot-constrained optimization objective balancing incremental net value against penalties for low-yield or non-persuadable interactions, including Always-Buyers and Never-Buyers. Rather than using LLMs as numerical causal calculators, CURA-CLM positions them as causal information organizers within a statistical reasoning pipeline.

The contributions of this paper are summarized as follows:

- We propose CURA-CLM, a modular LLM-assisted causal recommendation framework connecting unstructured textual evidence with formal statistical estimators, reducing the instability of direct LLM-based numerical causal inference.
- We formulate per-user recommendation allocation as a stratum-aware counterfactual reasoning problem under localized slot constraints, with a net-revenue-weighted objective that rewards incremental value while penalizing low-yield or non-persuadable interactions.
- Through evaluation on a synthetic dataset, we show that CURA-CLM improves over direct LLM recommendation, and uplift-ranking strategies with global constraints in preserving resources under strict per-user slot limitations.

2 Related Work

2.1 Large Language Models for Recommendation

Large language models have recently become an important tool for recommendation systems because they can encode high-density semantic evidence from user histories, review text, and item descriptions (Ren et al., 2024; Bao et al., 2023; Shi et al., 2025b; Yue et al., 2025; Wu et al., 2024; Shehmir & Kashef, 2025). Existing approaches use LLMs for user profiling, preference reasoning, semantic alignment, and open-ended recommendation generation. These efforts demonstrate that natural language context can improve recommendation quality beyond sparse ID-based interaction modeling. Nevertheless, the advantages of LLMs do not automatically translate into reliable stratum-aware allocation. A growing line of work shows that LLMs remain weak in formal causal reasoning, numerical inference, and counterfactual consistency (Jin et al., 2023; Liu et al., 2024b; Chen et al., 2023; Jin et al., 2024; Feder et al., 2022; Kiciman et al., 2023). This limitation is especially important when recommendation decisions are tied to scarce resource allocation. As a result, several recent studies have moved toward modular pipelines where LLMs serve as structured

information extractors rather than end-to-end decision makers. Text-to-table and multi-agent feature generation approaches such as COAT, Agent0, and RecXplore follow this principle (Liu et al., 2024a; Škrlj et al., 2024; Shi et al., 2025a). Our work adopts the same philosophy but places it in the concrete setting of recommendation slot allocation.

2.2 CATE Estimation and Constrained Treatment Allocation

The methodological basis of uplift modeling lies in the causal inference literature on heterogeneous treatment effects (Rubin, 1978; Rosenbaum & Rubin, 1983; Funk et al., 2005; Künzel et al., 2019; Athey et al., 2019; Nie & Wager, 2021; Kennedy et al., 2024). Methods such as X-Learner, DR-Learner, generalized random forests, and neural treatment-effect estimators have been widely studied for observational decision problems (Wager & Athey, 2018; Shi et al., 2019; Zhong et al., 2022). In recommendation settings, these methods provide a principled alternative to correlation-based ranking because they explicitly estimate the incremental effect of intervention.

At the same time, recommendation logs are affected by exposure bias, selection bias, popularity effects, and missing-not-at-random feedback (Schnabel et al., 2016; Wang et al., 2019; Wei et al., 2021; Li et al., 2023a; Gao et al., 2022; Luo et al., 2024; Zheng et al., 2026). Causal recommendation and constrained allocation have therefore become increasingly important. Existing work considers causal embeddings, disentangled recommendation, budget-constrained uplift recommendation, ROI-constrained promotion, and ranking under resource limitations (Bonner & Vasile, 2018; Zheng et al., 2021; Goldenberg et al., 2020; Sun et al., 2024; Kamran et al., 2024; Arno et al., 2026; De Vos et al., 2025; Wang et al., 2025; **?**; Liu et al., 2024c). However, prior work often emphasizes either causal estimation or constrained optimization in isolation, and many formulations remain centered on global expenditure rather than the per-user recommendation slot limitation that arises naturally in recommendation interfaces. The present paper aims to integrate text-derived covariates, treatment-effect estimation, and slot-constrained allocation within one coherent framework.

3 Problem Setup

Let $u \in \mathcal{U}$ denote a user and $i \in \mathcal{V}$ an item. For each user-item pair, a binary treatment variable $W_{u,i} \in \{0, 1\}$ indicates whether item i is recommended to user u, and a binary outcome variable $Y_{u,i} \in \{0, 1\}$ indicates whether the user purchases the item. Following the potential-outcomes framework (Rubin, 1978; Hernán & Robins, 2020), we define two potential outcomes, $Y_{u,i}(1)$ and $Y_{u,i}(0)$, corresponding to purchase with and without recommendation. Given a structured covariate vector $X_{u,i}$ extracted from text, the conditional average treatment effect is

$$\tau(X_{u,i}) = \mathbb{E}[Y_{u,i}(1) - Y_{u,i}(0) \mid X_{u,i}]. \tag{1}$$

We assume that the observed outcome exactly matches the potential outcome under the realized treatment, expressed as $Y_{u,i} = W_{u,i}Y_{u,i}(1) + (1 - W_{u,i})Y_{u,i}(0)$, which is known as the consistency assumption in the causal framework. We also assume that the treatment assigned to one user-item pair does not interfere with or alter the potential outcomes of any other pair, ensuring that the Stable Unit Treatment Value Assumption (SUTVA) holds. Furthermore, we assume $(Y_{u,i}(1), Y_{u,i}(0)) \perp\!\!\!\perp W_{u,i} \mid X_{u,i}$, meaning the treatment assignment is conditionally independent of the potential outcomes once the structured covariates are observed; this is formally referred to as unconfoundedness. Finally, we assume $0 < \mathbb{P}(W_{u,i} = 1 \mid X_{u,i}) < 1$, ensuring both treatment states remain observable for all relevant covariate patterns, which is the positivity assumption. Under these conditions, heterogeneous treatment effects can be identified from observational data and estimated with standard causal meta-learners.

This setup clarifies why per-user-slot-constrained recommendation differs from standard ranking with global budget. A user-item pair is an inefficient intervention target if its untreated baseline is already high; what truly matters is the incremental change in purchase probability. While intuitively related to counterfactual groups like persuadables or always-buyers, our optimization avoids an explicit response-pattern taxonomy (Li et al., 2023b). Instead, we introduce two user-item economic parameters: $v_{u,i}$, the net revenue from a successful incremental conversion, and $c_{u,i}$, the platform-side cost (e.g., promotional or opportunity cost) of assigning a slot. Accordingly, the recommendation list acts as a constrained treatment allocation whose utility depends jointly on the estimated uplift, the conversion value, and the slot cost.

4 METHOD

To address the waste-of-slots problem under localized per-user capacity constraints, we propose CURA-CLM, an end-to-end causal recommendation framework. The framework contains three components. First, it extracts structured causal covariates from unstructured user and item text by using LLMs as semantic parsers over a computation-manageable candidate set. Second, it applies robust statistical estimators to these structured features to identify the Conditional Average Treatment Effect (CATE). Third, it translates the estimated causal effects into recommendation decisions through a per-user-slot-constrained optimization objective that weights uplift by the potential net revenue of a successful conversion while penalizing the cost of allocating recommendation slots to low-yield or invalid user-item interactions, such as never-/always-buyers.

4.1 CANDIDATE CONSTRUCTION AND TEXT-TO-TABLE CONVERSION

The first stage resolves the computational mismatch between large item catalogs and costly LLM processing. Rather than evaluating the full universe, a lightweight retrieval step constructs a manageable candidate set per user. This step does not solve the causal allocation problem itself, but efficiently narrows the pool for downstream scoring.For each retained pair, a text-to-table module bridges unstructured text and the structured inputs required by causal estimators, adapting the Causal representatiOn AssistanT (COAT) paradigm (Liu et al., 2024a). Let T_u^{user} be the user's textual history and T_i^{item} the item description. An LLM operator Ψ with prompt $\mathcal{P}$ produces a structured covariate vector:

$$X_{u,i} = \Psi(T_u^{\text{user}}, T_i^{\text{item}}, \mathcal{P}). \tag{2}$$

Following COAT, extraction occurs in two phases: factor proposal and annotation. The LLM first identifies confounding causal factors (e.g., category preference, affordability, promotion responsiveness). It then assigns discrete values, yielding $X_{u,i} \in \{-1, 0, 1\}^d$, where $+1$ indicates the factor increases purchase likelihood upon recommendation, -1 indicates a decrease, and 0 is neutral.By restricting the LLM strictly to semantic extraction, this design enables downstream uplift estimation on structured covariates while avoiding the unstable numerical hallucinations associated with end-to-end LLM causal inference (Jin et al., 2023; Liu et al., 2024b). This proves far more reliable than prompting models directly for treatment-effect estimation or allocation decisions.

4.2 TREATMENT-EFFECT ESTIMATION WITH X-LEARNER AND DR-LEARNER

After obtaining the structured covariates, we estimate heterogeneous treatment effects using X-Learner and DR-Learner.

X-Learner is well suited to imbalanced treatment and control samples (Künzel et al., 2019). Let $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$ denote the outcome regressors fitted on treated and untreated data. Pseudo-effects are first imputed for the two groups:

$$\hat{D}_{u,i}^{(1)} = Y_{u,i} - \hat{\mu}_0(X_{u,i}), \quad \text{for } W_{u,i} = 1, \quad \hat{D}_{u,i}^{(0)} = \hat{\mu}_1(X_{u,i}) - Y_{u,i}, \quad \text{for } W_{u,i} = 0. \tag{3}$$

Then two effect models $\hat{\tau}_1(X)$ and $\hat{\tau}_0(X)$ are learned from the treated and untreated pseudo-effects, respectively. The final treatment-effect estimate combines them through the propensity score:

$$\hat{\tau}_X(X_{u,i}) = \hat{e}(X_{u,i})\hat{\tau}_0(X_{u,i}) + \big(1 - \hat{e}(X_{u,i})\big)\hat{\tau}_1(X_{u,i}), \tag{4}$$

where $\hat{e}(X_{u,i})$ denotes the estimated propensity of recommendation.

DR-Learner uses a doubly robust pseudo-outcome based on both an outcome model and a propensity model (Funk et al., 2005). Let $\hat{e}(X_{u,i})$ denote the propensity score and $\hat{\mu}_t(X_{u,i})$ the predicted outcome under treatment state $t \in \{0, 1\}$. The pseudo-outcome is

$$\hat{Y}_{u,i}^{DR} = \left(\frac{W_{u,i}(Y_{u,i} - \hat{\mu}_1(X_{u,i}))}{\hat{e}(X_{u,i})} + \hat{\mu}_1(X_{u,i}) \right) \tag{5}$$

$$- \left(\frac{(1 - W_{u,i})(Y_{u,i} - \hat{\mu}_0(X_{u,i}))}{1 - \hat{e}(X_{u,i})} + \hat{\mu}_0(X_{u,i}) \right). \tag{6}$$

A regression model on $\hat{Y}_{u,i}^{DR}$ yields the final estimate $\hat{\tau}_{DR}(X_{u,i})$. The attraction of DR-Learner in observational recommendation settings is that consistency is retained if either the response model or the propensity model is well specified.

4.3 NET-REVENUE-WEIGHTED SLOT-CONSTRAINED ALLOCATION

The final component converts treatment-effect estimates into recommendation decisions under per-user recommendation slot limitation. Let $D_{u,i} \in \{0,1\}$ indicate whether item i is selected for user u, and let K denote the maximum number of recommendation slots available to each user. CURA-CLM solves the following slot-constrained allocation problem:

$$\max_{D_{u,i}} \sum_{i \in \mathcal{C}_u} D_{u,i} \hat{r}_{u,i} \quad \text{s.t.} \quad \sum_{i \in \mathcal{C}_u} D_{u,i} \leq K, \quad D_{u,i} \in \{0,1\}, \tag{7}$$

where $\mathcal{C}_u$ is the computation-manageable candidate set for user u, and $\hat{r}_{u,i}$ is the estimated reward score used for both optimization and ranking. This formulation makes the per-user slot limitation explicit: among all candidate items, only K can be allocated to the final recommendation list.

Let $r_{u,i}$ denote the ideal reward score of recommending item i to user u. We define it as

$$r_{u,i} = v_{u,i}\tau(X_{u,i}) - \lambda c_{u,i}\big(1 - \tau(X_{u,i})\big), \tag{8}$$

where $v_{u,i}$ denotes the potential net revenue gain from one successful incremental conversion, $c_{u,i}$ denotes the cost incurred if the platform insists on allocating a recommendation slot to this user-item pair, and $\lambda > 0$ controls the strength of the cost penalty. The first term weights the uplift effect by the potential net revenue gain of a successful conversion. The second term penalizes invalid or low-yield allocation: for example, recommending to always-buyers wastes a slot because the purchase would likely occur anyway, while recommending to never-buyers consumes slot resources without producing conversion gain.

Since the true CATE $\tau(X_{u,i})$ is unobserved, the reward score must be estimated from the learned treatment-effect model. The reward score estimator is

$$\hat{r}_{u,i} = v_{u,i}\hat{\tau}(X_{u,i}) - \lambda c_{u,i}\big(1 - \hat{\tau}(X_{u,i})\big). \tag{9}$$

In our implementation, λ is tuned on a validation split and fixed to $\lambda = 0.3$.

The final recommendation list is obtained by sorting the candidate items according to $\hat{r}_{u,i}$ and selecting the top K for each user. Because each item consumes one slot, this greedy Top-K solution is optimal under the unit-slot assumption. The key point is that the selection rule is no longer driven by relevance alone or by uplift alone, but by a slot-aware estimate of intervention utility.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Since real observational logs cannot simultaneously reveal both potential outcomes (Rubin, 1978; Hernán & Robins, 2020), we follow standard practice (Shalit et al., 2017; Lyu et al., 2022) by evaluating CURA-CLM on a synthetic benchmark with ground-truth effects. The dataset, comprising 500 users, 100 items, and 10,000 interactions, is generated in four steps. First, we construct diverse user and item pools. Second, we assign each pair a 12-dimensional covariate vector $X_{u,i} \in \{-1, 0, 1\}^{12}$ capturing causal factors (e.g., price sensitivity, urgency). Third, a structural causal model generates the baseline outcome $Y_{u,i}(0) \sim \text{Bernoulli}(\sigma(\alpha_0 + \sum_{k=1}^{12} \alpha_k X_{u,i,k}))$ and true incremental effect $\tau_{u,i} = \sigma(\gamma_0 + \sum_{k=1}^{12} \gamma_k X_{u,i,k})$, yielding $Y_{u,i}(1) \sim \text{Bernoulli}(Y_{u,i}(0) + \tau_{u,i})$, storing $\tau_{u,i}$ for explicit evaluation. Finally, LLMs synthesize user and item texts that naturally embed these covariates, ensuring our text-to-table module is rigorously evaluated on realistic semantic inputs rather than oracle structured features.

Evaluation Metrics. We report six metrics aligned with recommendation slot economics: GMV, Relative GMV, Net Revenue, Relative Net Revenue, Positives, and Positive Rate.

Baselines. We compare CURA-CLM with three families of baselines: (1) Random as an uninformed allocation reference; (2) Text-based CF, a correlation-oriented retrieval baseline built on semantic similarity in the spirit of text-driven recommendation models (Ren et al., 2024; Shi et al., 2025b); (3) Vanilla-CoT, a direct LLM recommendation baseline motivated by chain-of-thought prompting (Wei et al., 2022; Yue et al., 2025); and (4) Top-K Uplift and Top-K GMV Incremental, structured causal-ranking baselines derived from standard treatment-effect estimation and constrained ranking ideas (Künzel et al., 2019; Funk et al., 2005; Kamran et al., 2024; Sun et al., 2024).

Table 1: Performance comparison under per-user recommendation slot limitation $K = 10$ using GPT-4o and Qwen3.5-plus. Bold indicates the best results for each LLM backbone.

Backbone	Method	Estimator	GMV	Rel. GMV	Net Revenue	Rel. Revenue	Positives	Positive Rate
GPT-4o	Random	–	4,320,966.18	0.0000	-160,399.95	0.0000	2,890	0.4951
	Text-based CF	–	5,430,386.86	0.2568	-451,090.45	1.8123	2,043	0.3497
	Vanilla-CoT	–	4,440,688.84	-0.0033	1,831,455.21	-0.0793	2,846	0.4875
	Top-K Uplift	X-Learner	3,511,692.97	-0.1873	1,698,201.25	-0.0821	**3,358**	**0.5755**
	Top-K Uplift	DR-Learner	3,646,913.14	-0.1560	1,739,399.04	-0.0598	3,312	0.5670
	Top-K GMV Inc.	X-Learner	5,394,933.24	0.2485	2,548,239.26	0.3774	3,130	0.5363
	Top-K GMV Inc.	DR-Learner	5,807,340.54	0.3440	2,637,537.50	0.4257	3,029	0.5189
	CURA-CLM	X-Learner	7,168,158.95	0.6589	**3,249,179.74**	**0.7563**	2,930	0.5021
	CURA-CLM	DR-Learner	**7,192,316.51**	**0.6645**	3,237,583.98	0.7500	2,952	0.5059
Qwen3.5-plus	Random	–	4,320,966.18	0.0000	-160,399.95	0.0000	2,890	0.4951
	Text-based CF	–	5,430,386.86	0.2568	-451,090.45	1.8123	2,043	0.3497
	Vanilla-CoT	–	4,740,234.44	0.0639	2,036,181.98	0.0236	2,761	0.4731
	Top-K Uplift	X-Learner	3,539,639.41	-0.1808	1,664,888.49	-0.1001	**3,355**	**0.5749**
	Top-K Uplift	DR-Learner	3,627,290.51	-0.1605	1,734,851.77	-0.0623	3,321	0.5688
	Top-K GMV Inc.	X-Learner	5,143,639.84	0.1904	2,460,344.88	0.3299	3,151	0.5399
	Top-K GMV Inc.	DR-Learner	5,384,809.12	0.2462	2,508,926.56	0.3561	3,105	0.5320
	CURA-CLM	X-Learner	6,688,745.95	0.5480	2,993,550.43	0.6181	3,014	0.5163
	CURA-CLM	DR-Learner	**6,788,406.67**	**0.5710**	**3,060,256.41**	**0.6542**	3,028	0.5187

Implementation Details. We evaluate two LLM backbones for the text-to-table stage, GPT-4o (OpenAI, 2024) and Qwen3.5-plus (Team, 2026). The candidate set size is fixed to 15 items per user, and $K = 10$ for per-user recommendation slots. Text-to-table extraction is performed with deterministic prompting and low-temperature decoding. Causal estimation is implemented with DoWhy v0.14 and EconML v0.15, with propensity scores estimated through regularized logistic regression and clipped to $[0.05, 0.95]$. The X-Learner and DR-Learner base regressors are instantiated with gradient-boosting style models. The penalty coefficient is fixed to $\lambda = 0.3$ after validation. All experiments are conducted on a server with an NVIDIA A40 GPU and 128GB RAM.

5.2 Performance Comparison

We compare overall allocation performance under the strict per-user recommendation slot limitation of $K = 10$. Since the text-to-table stage can be instantiated with different LLM backbones, we report results for GPT-4o and Qwen3.5-plus in a single merged presentation table. This experiment evaluates whether the proposed framework improves recommendation allocation under per-user slot constraint and whether the combination of text-to-table conversion, causal estimation, and stratum-aware allocation yields consistent gains over correlation-based retrieval, direct LLM recommendation, and simpler uplift-ranking strategies. Table 1 yields several key observations. First, all causal methods, especially CURA-CLM and Top-K GMV Incremental, consistently outperform the Random, Text-based CF, and Vanilla-CoT baselines across both LLM backbones, validating the advantage of text-to-table conversion and treatment-effect estimation over simple correlational similarity. Second, while Top-K Uplift yields a higher Positive Rate, these raw counts are misleading; allocating slots to low-yield or invalid groups severely degrades actual GMV and Net Revenue, exemplifying the waste-of-slots problem. Finally, although Top-K GMV Incremental accounts for item value, it ignores the heterogeneous resource impact across response strata. By explicitly penalizing these invalid strata, CURA-CLM successfully filters out wasteful allocations, maximizing net revenue and overall slot allocation quality under strict per-user constraints.

6 Conclusion

This paper addresses the "waste-of-slots" problem in text-driven recommendation systems under strict per-user capacity limits. We introduce CURA-CLM, an end-to-end framework that integrates LLM-based semantic-to-structured causal abstraction, robust treatment-effect estimation, and stratum-aware allocation. By penalizing slot assignments to non-persuadable interactions, CURA-CLM significantly outperforms direct LLM prompting, and standard uplift ranking in maximizing net revenue. More broadly, our results support a modular reasoning paradigm: LLMs should organize textual evidence into causally relevant representations, while numerical causal estimation and constrained allocation are handled by formal statistical models. Future work may extend this framework to hidden confounding and dynamic cost structures in real-world deployments.

REFERENCES

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–472, 1996.

Henri Arno, Dennis Frauen, Emil Javurek, Thomas Demeester, and Stefan Feuerriegel. Ranklearner: Orthogonal ranking of treatment effects. *arXiv preprint arXiv:2602.03517*, 2026.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 85(1):1–26, 2017.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Keqin Bao, Jizhi Zhang, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. Tallrec: An effective and efficient tuning framework to align large language model with recommendation. In *ACM Conference on Recommender Systems*, 2023. arXiv:2305.00447.

Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *ACM Conference on Recommender Systems*, pp. 104–112, 2018.

Wenhu Chen, Xuezhi Ma, Xuezhi Wang, Michael Lyu, James Hong, William Wang, Jie Liu, Bo Zhang, Jing Shi, and Zhaopeng Tu. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *TMLR*, 2023.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. Wide and deep learning for recommender systems. In *Workshop on Deep Learning for Recommender Systems*, 2016.

Simon De Vos, Jente Van Belle, Andres Algaba, Wouter Verbeke, and Sam Verboven. Decision-centric fairness: Evaluation and optimization for resource allocation problems. *arXiv preprint arXiv:2504.20642*, 2025.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 61(4): 962–973, 2005.

Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems*, 41(3):1–32, 2022.

Dmitri Goldenberg, Javier Albert, Lucas Bernardi, and Pablo Estevez. Free lunch! retrospective uplift modeling for dynamic promotions recommendation within roi constraints. In *ACM Conference on Recommender Systems*, 2020.

Dmitri Goldenberg, Amit Livne, Carlos Herrero-Gómez, Javier Albert, Itsik Adiv, Igor Spivak, Ran Dach, Felipe Moraes, Hugo Manuel Proença, and Bracha Shapira. Challenges and methods of causal promotions recommendation in e-commerce. *ACM Transactions on Information Systems*, 2025.

Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, Xiuqiang He, and Zhenhua Dong. Deepfm: An end-to-end wide & deep learning framework for ctr prediction. In *The International Joint Conference on Artificial Intelligence*, 2018.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *International World Wide Web Conference*, 2017.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. CRC Press, 2020.

Jianqing Jiang, Madya Nor Asilah Wati Abdul Hamid, Ng Keng Yap, and Madya Choo Wei Chong. A dynamic framework for causal user profiling and treatment segmentation via uplift modeling in internet lending. *IEEE Access*, 2026.

Zhijing Jin, Yanda Chen, Felix Leeb, Omar Kamal, Zhiheng Lyu, Kai Blin, Francisco Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models. In *Conference on Neural Information Processing Systems*, 2023.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *The International Conference on Learning Representations*, 2024.

Fahad Kamran, Maggie Makar, and Jenna Wiens. Learning to rank for optimal treatment allocation under resource constraints. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Edward H Kennedy, Sivaraman Balakrishnan, James M Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. *Annals of statistics*, 52(2):793–823, 2024.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2023.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, Peng Wu, and Peng Cui. Propensity matters: Measuring and enhancing balancing for recommendation. In *International Conference on Machine Learning*, 2023a.

Haoxuan Li, Chunyuan Zheng, Peng Wu, Kun Kuang, Yue Liu, and Peng Cui. Who should be given incentives? counterfactual optimal treatment regimes learning for recommendation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023b.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. Discovering and reasoning of causality in the hidden world with large language models. In *Conference on Neural Information Processing Systems*, 2024a.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics*, pp. 9215–9235, 2024b.

Xu Liu, Tong Yu, Kaige Xie, Junda Wu, and ShThe Conference on Uncertainty in Artificial Intelligence Li. Interact with the explanations: Causal debiased explainable recommendation system. In *ACM Web Search and Data Mining*, 2024c.

Huishi Luo, Fuzhen Zhuang, Ruobing Xie, Hengshu Zhu, Deqing Wang, Zhulin An, and Yongjun Xu. A survey on causal inference for recommendation. *The Innovation*, 5(2), 2024.

Yan Lyu, Sunhao Dai, Peng Wu, Quanyu Dai, Yuhao Deng, Wenjie Hu, Zhenhua Dong, Jun Xu, Shengyu Zhu, and Xiao-Hua Zhou. A semi-synthetic dataset generation framework for causal inference in recommender systems. *arXiv preprint arXiv:2202.11351*, 2022.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation learning with large language models for recommendation. In *International World Wide Web Conference*, pp. 3874–3884, 2024.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *The Conference on Uncertainty in Artificial Intelligence*, 2009.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

Yuta Saito, Hayato Sakata, and Kazuhide Nakata. Doubly robust prediction and evaluation methods improve uplift modeling for observational data. In *SIAM International Conference on Data Mining*, 2019.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *International Conference on Machine Learning*, pp. 1670–1679, 2016.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *International Conference on Machine Learning*, 2017.

Sarama Shehmir and Rasha Kashef. Llm4rec: a comprehensive survey on the integration of large language models in recommender systems—approaches, applications and challenges. *Future Internet*, 17(6):252, 2025.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Conference on Neural Information Processing Systems*, 2019.

Kainan Shi, Peilin Zhou, Ge Wang, Han Ding, and Fei Wang. What matters in llm-based feature extractor for recommender? a systematic analysis of prompts, models, and adaptation. *arXiv preprint arXiv:2509.14979*, 2025a.

Xiaofeng Shi, Yuduo Li, Qian Kou, Longbin Yu, Jinxin Xie, and Hua Zhou. Spar: Scholar paper retrieval with llm-based agents for enhanced academic search. *arXiv preprint arXiv:2507.15245*, 2025b.

Blaž Škrlj, Benoît Guilleminot, and Andraž Tori. Agent0: Leveraging llm agents to discover multivalue features from text for enhanced recommendations. In *Agent4IR*, 2024.

Zexu Sun, Hao Yang, Dugang Liu, Yunpeng Weng, Xing Tang, and Xiuqiang He. End-to-end cost-effective incentive recommendation under budget constraint with uplift modeling. In *ACM Conference on Recommender Systems*, 2024.

Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Hao Wang, Zhichao Chen, Honglei Zhang, Zhengnan Li, Licheng Pan, Haoxuan Li, and Mingming Gong. Debiased recommendation via wasserstein causal balancing. *ACM Transactions on Information Systems*, 43(6):1–24, 2025.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pp. 6258–6267, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Conference on Neural Information Processing Systems*, 2022.

Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1927–1937, 2021.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, 2024.

Weiqi Yue, Yuyu Yin, Xin Zhang, Binbin Shi, Tingting Liang, and Jian Wan. Cot4rec: Revealing user preferences through chain of thought for recommender systems. In *The AAAI Conference on Artificial Intelligence*, 2025.

Chunyuan Zheng, Haocheng Yang, Jinkun Chen, Shufeng Zhang, and Tianyu Xia. Unified minimax optimization framework for propensity score estimation in debiased recommendation. In *The AAAI Conference on Artificial Intelligence*, 2026.

Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *International World Wide Web Conference*, pp. 2988–2999, 2021.

Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. Descn: Deep entire space cross networks for individual treatment effect estimation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.