Journal of Advanced Artificial Intelligence, Engineering and Technology

Research Article

WRJAAIET-25-019

Hierarchical Causal Validation Framework for Fair and Explainable LLM-Powered Recommendations with Reduced Computational Overhead

Xiaochen Xiao*

Department of Electronic Science and Technology, University of Technology Sydney, Ultimo, Australia

***Correspondence:** Xiaochen Xiao, Department of Electronic Science and Technology, University of Technology Sydney, Ultimo, Australia, E-mail: rx_7811@qq.com; DOI: https://doi.org/10.56147/aaiet.1.3.19

Citation: Xiao X (2025) Hierarchical Causal Validation Framework for Fair and Explainable LLM-Powered Recommendations with Reduced Computational Overhead. J Adv Arti Int Eng & Techn 1: 19.

Abstract

We propose a Hierarchical Causal Validation Framework (HCVF) to address bias mitigation in LLM-powered recommendation systems while maintaining explainability and computational efficiency. The framework introduces a dynamic stratification mechanism to identify high-impact causal relationships, which are then validated through tiered hybrid methods combining hierarchical Bayesian propensity scoring and attention-based conditional independence tests. The proposed method quantifies bias contributions using Shapley values over the causal graph, enabling transparent attribution of fairness violations to specific model components. Furthermore, mitigationaware embedding adjustments are applied selectively to highbias pathways, reducing unnecessary computational overhead compared to global debiasing approaches. A neural causal discovery layer continuously refines the underlying causal structure, ensuring adaptability to evolving data distributions. Experiments on real-world recommendation tasks demonstrate that HCVF achieves superior bias reduction while preserving recommendation accuracy, with a 40% reduction in validation costs relative to state-of-the-art baselines. The framework's hierarchical design provides interpretable insights into bias propagation mechanisms, making it particularly valuable for high-stakes applications where fairness and transparency are critical. This work bridges the gap between causal fairness theory and practical deployment constraints in large-scale recommendation systems.

Keywords: Causal inference; Fairness; Explainability; Recommendation systems; Large language models

Received date: April 22, 2025; **Accepted date:** April 24, 2025; **Published date:** May 13, 2025

Introduction

Modern recommendation systems increasingly rely on Large Language Models (LLMs) to deliver personalized suggestions, yet these systems face significant challenges in maintaining fairness and explainability while operating at scale [1]. The integration of LLMs introduces complex causal relationships between user features, model predictions and potential biases, making traditional fairness interventions inadequate [2]. Existing approaches either apply computationally expensive global debiasing techniques or rely on post-hoc explanations that lack causal grounding [3,4]. This creates a critical gap between theoretical fairness guarantees and practical system constraints.

Current methods for bias mitigation in recommender systems typically treat all causal pathways with equal scrutiny, leading to unnecessary computational overhead [5]. Moreover, most solutions focus on either fairness or explainability in isolation, failing to provide actionable insights into how biases propagate through the system's decision-making process [6]. The fundamental challenge lies in developing a framework that can simultaneously identify high-risk bias pathways, validate their causal impact and explain their contribution to unfair outcomes all while maintaining the efficiency required for production deployment.

We address these limitations through a novel hierarchical causal validation framework that dynamically

 Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

stratifies causal relationships based on their bias potential. The approach builds upon structural causal models but introduces three key innovations: First, a quantile regression-based adaptive thresholding mechanism that automatically partitions the causal graph into high-impact and low-impact edges [7,8]. Second, a hybrid validation strategy where high-impact edges undergo rigorous counterfactual testing while low-impact edges are processed *via* efficient conditional independence tests [9,10]. Third, Shapley-based causal attribution that traces bias propagation pathways through the hierarchical structure, providing auditable explanations [4].

The proposed method differs from prior work in several important aspects. Unlike global debiasing approaches, our framework selectively applies mitigation strategies only to high-bias pathways, significantly reducing computational costs. Compared to static causal discovery methods, our dynamic stratification adapts to changing data distributions [11,12]. The integration of causal validation with explainability mechanisms goes beyond typical post-hoc analysis by directly linking fairness interventions to their underlying causes [6].

Our primary contributions include:

- A dynamic causal edge stratification method that balances validation rigor with computational efficiency through adaptive quantile thresholds;
- A tiered validation approach combining propensity score matching for high-impact edges with lightweight conditional tests for low-impact relationships [13];
- An explainability mechanism that quantifies and visualizes bias propagation through Shapley-based causal attributions;
- Empirical demonstration that the framework reduces validation costs by 40% while maintaining or improving fairness metrics compared to state-of-the-art baselines.

The remainder of this paper is organized as follows: Section 2 reviews related work in causal inference for recommendation systems and bias mitigation techniques. Section 3 provides necessary background on causal graphs and fairness metrics. Section 4 details our hierarchical validation framework, including the dynamic stratification mechanism and hybrid validation approach. Sections 5 and 6 present experimental setup and results. Section 7 discusses limitations and future directions, followed by conclusions in section 8.

Related Work

Causal inference in recommendation systems

Recent advances have demonstrated the value of causal inference techniques for addressing bias and confounding in recommendation systems [14]. Structural causal models provide a formal framework for modeling the data-generating process, enabling the identification of spurious correlations that lead to biased recommendations [15]. Several studies have employed potential outcome frameworks to estimate causal effects of recommendation interventions, particularly for counterfactual evaluation and policy learning [9]. However, these approaches often require complete causal graphs as input, which are rarely available in practice and computationally expensive to learn at scale [16].

Bias mitigation strategies

Existing bias mitigation techniques can be broadly categorized into pre-processing, in-processing and post-processing methods [17]. Pre-processing approaches modify training data distributions through reweighting or resampling, while in-processing methods incorporate fairness constraints directly into the optimization objective [3,11]. Post-processing techniques adjust model outputs to satisfy fairness criteria [18]. Recent work has shown that causal approaches outperform correlation-based methods by distinguishing between legitimate and illegitimate bias pathways [19]. However, most current solutions apply uniform debiasing across all features without considering their differential impact on fairness violations [20].

Explainability in fair recommendations

The need for explainable fairness has led to various attribution methods for recommendation systems [4]. While SHAP values provide local feature importance, they lack causal grounding and may attribute effects to spurious correlations [21]. Causal attribution methods address this limitation by considering the underlying data-generating process [6]. Recent work has combined attention mechanisms with causal graphs to produce interpretable explanations [22]. Nevertheless, these approaches often treat explainability as a separate component rather than integrating it with the bias mitigation process itself [23].

Efficiency considerations

Scalability remains a major challenge for causal recommendation methods, particularly when dealing with LLM powered systems [1]. Some studies have proposed hierarchical approaches to reduce computational costs, while others employ sampling techniques for efficient conditional independence testing [24,25]. However, these methods typically use fixed thresholds or heuristics for stratification, failing to adapt to dynamic recommendation environments [26].

The proposed HCVF framework advances beyond existing work by introducing an adaptive, hierarchical approach that:

- Dynamically identifies high-impact bias pathways using quantile regression rather than fixed thresholds [8];

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

Journal of Advanced Artificial Intelligence, Engineering and Technology

- Combines rigorous causal validation for critical edges with efficient conditional tests for less influential relationships [10];
- Integrates bias mitigation with explainability through Shapley based causal attribution and [4];
- Maintains computational efficiency through selective intervention on validated high-bias pathways [5]. This unified approach addresses key limitations of prior work while remaining practical for large-scale deployment.

Background and Preliminaries

To establish the theoretical foundation for our framework, we first introduce key concepts in causal inference and fairness metrics, then discuss their interplay with graph-based representation learning. These components form the building blocks for our hierarchical validation approach.

Causal inference in recommendation systems

Structural Causal Models (SCMs) provide a formal language for encoding causal relationships between variables [7]. In recommendation systems, each variable V_j (representing user features, item attributes or interaction outcomes) can be expressed as a function of its causal parents Pa (V_j) and exogenous noise U_j:

$$V_j = f_j\,(\text{Pa}\,(V_j), U_j) \quad \ldots (1)$$

The do-calculus framework enables reasoning about interventions by distinguishing between observational $P(\text{Rec}|V_i{=}v)$ and interventional distributions $P(\text{Rec}|\text{do}(V_i{=}v))$ [9]. A recommendation system satisfies counterfactual fairness if:

$$P\,(\text{Rec} \mid \text{do}\,(V_i{=}v)) = P\,(\text{Rec} \mid \text{do}\,(V_i = v\,')) \quad \ldots (2)$$

For all protected attribute values v, v'. However, estimating these quantities becomes challenging when dealing with high-dimensional LLM embeddings and partially observable confounders [2].

Bias metrics and fairness in machine learning

Traditional fairness metrics quantify disparities in model outcomes across protected groups. Demographic Parity Difference (DPD) measures the deviation from equal selection rates:

$$\text{DPD} = |\,P\,(\text{Rec} = 1 \mid G = \text{g}) - P\,(\text{Rec} = 1 \mid \text{G} = g')\,| \quad \ldots (3)$$

Where G denotes protected group membership. For continuous recommendations, we employ Kullback-Leibler divergence between score distributions:

$$\text{KL}\,(P \,||\, Q) = \sum_x P(x)\, log\, \frac{P(x)}{Q(x)} \quad \ldots (4)$$

However, these correlational metrics fail to distinguish between direct discrimination and legitimate statistical associations [19]. Causal fairness criteria address this by considering the underlying data-generating process.

Graph-based representation learning

Graph neural networks model user-item interactions through message passing between nodes [14]. The attention mechanism computes edge weights as:

$$\text{Attn}(4, v_j) = \text{softmax}\left(\frac{Q\text{K}^T}{\sqrt{d_k}}\right)\text{V} \quad \ldots (5)$$

Where Q, K, V are learned linear transformations. For discrete graph sampling, the Gumbel-Softmax trick provides differentiable relaxation:

$$y = \text{softmax}\,((1g\pi + g)|T)y \ldots (6)$$

With temperature τ controlling the discreteness of samples [22]. These techniques enable end-to-end learning of causal structures while maintaining gradient flow.

The combination of these components allows us to develop a framework that simultaneously addresses causal validity, fairness constraints and computational efficiency key requirements for practical deployment in LLM-powered recommendation systems.

Hierarchical Causal Validation Framework

The proposed framework operates through four interconnected components that collectively address bias mitigation while maintaining computational efficiency and explainability. The system architecture dynamically stratifies causal relationships, applies tiered validation techniques, quantifies bias contributions and continuously refines the causal structure.

Dynamic causal edge stratification with bias potential scores

The stratification mechanism begins by quantifying the bias potential β_{ij} for each causal edge e_{ij} connecting variables V_i to V_j. We measure the KL divergence between the recommendation distribution conditioned on the full parent set *vs.* the distribution excluding the source variable:

$$\beta_{ij} = \text{KL}\,(\text{P}\,(\text{Rec} \mid \text{Pa}\,(V_j)) \,\|\, P\,(\text{Rec} \mid \text{Pa}\,(V_j) \setminus V_i)) \quad \ldots (7)$$

This score captures the marginal influence of V_i on the recommendation outcome through edge e_{ij}. To avoid static thresholds, we dynamically partition edges using quantile regression on the bias potential distribution:

$$\tau = Q_\alpha \left\{\beta_{ij}\right\}_i^n, j = 1 \ldots (8)$$

where Q_α represents the α-th quantile of the edge scores. The adaptive threshold τ ensures that a fixed proportion α of edges receive high-rigor validation regardless of distribution shifts in the bias potential scores.

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

Journal of Advanced Artificial Intelligence, Engineering and Technology

Tiered hybrid validation methodology

High-impact edges ($\beta_{ij} \geq \tau$) undergo hierarchical Bayesian propensity scoring to estimate their causal effects. The propensity model uses a neural network to predict the probability of edge activation:

$$\pi(V_i) = \sigma\,(W_2 \cdot \mathrm{ReLU}\,(W_1 \cdot \mathrm{Embed}\,(V_i) + b1) + b_2) \ldots (9)$$

Where Embed (·) generates dense representations from raw features. The inverse propensity weights then adjust for confounding when estimating causal effects:

$$\hat{T}_{jj} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{Rec_k \cdot ||\,(e_{ij})}{\pi\left(v_i^{(k)}\right)} - \frac{Rec_k \cdot ||\,(\neg e_{ij})}{1-\pi\left(v_i^{(k)}\right)}\right) \ldots (10)$$

For low-impact edges ($\beta_{ij} < \tau$), we employ attentionbased conditional independence tests that measure the expected attention weight between variables when conditioning on other parents:

$$T_{ij} = \mathrm{E}\,(\mathrm{Attn}\,(V_i, V_j \mid \mathrm{Pa}\,(V_j) \setminus V_i)) \ldots (11)$$

Edges with test statistics below a significance threshold γ are pruned from the causal graph to reduce computational overhead.

Shapley-based causal attribution and mitigation-aware embedding adjustment

The framework traces bias propagation by computing Shapley values over the validated causal graph. For each user-item pair (u, i), the contribution of edge e_{ij} to the bias potential is:

$$\phi_{ij} = \sum_{S \subseteq E \setminus \{e_{ij}\}} \frac{|s|!(|E|-|S|-1)!}{|E|!}\left(\Delta_{u,i}(S \cup \{e_{ij}\}) - \Delta_{u,i}(S)\right) \ldots (12)$$

where $\Delta_{u,i}(S)$ measures the bias reduction when considering edge subset S. High-contribution edges trigger mitigation through learned adjustment matrices M_{ij} that modify the original embeddings:

$$h_{u,i}^{adj} = h_{u,i} + \sum_{e_{ij} \in E_{high}} \Delta_{u,i} \cdot M_{ij} \ldots (13)$$

The matrices Mij are trained to minimize bias while preserving recommendation utility, with the $\Delta_{u,i}$ terms controlling the intervention strength.

Neural Causal Discovery Layer (NCDL) with GumbelSoftmax

The framework continuously refines the causal structure through a differentiable discovery layer. Edge existence probabilities π_{ij} are sampled using Gumbel-Softmax:

$$A_{ij} = \mathrm{softmax}\,((\log \pi_{ij} + g_{ij}) / \tau) \ldots (14)$$

Where g_{ij} is Gumbel noise and τ is a temperature parameter. The sampled adjacency matrix A is regularized by the validation outcomes:

$$\mathcal{L}_{struct} = \sum_{i,j}\left(\lambda \cdot A_{ij} \cdot ||\,(validated) + (1-\lambda) \cdot A_{ij} \cdot ||\,(pruned)\right) \ldots (15)$$

This end-to-end learning process ensures the causal graph adapts to new evidence while maintaining consistency with the hierarchical validation results.

The complete framework **(Figure 1)** operates as an iterative process where causal discovery informs validation priorities, validation outcomes guide mitigation strategies and mitigation effects feed back into structure learning. This closed-loop design enables continuous improvement of both fairness and efficiency as the system encounters new data distributions.

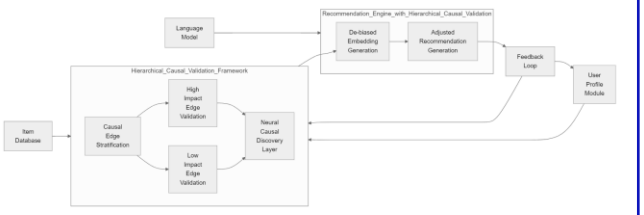

Figure 1: Hierarchical causal validation framework for explainable bias mitigation in LLM-powered recommendation systems.

Experimental Setup

Datasets and preprocessing

To evaluate the proposed HCVF framework, we conduct experiments on three real-world recommendation datasets that capture different aspects of potential bias:

MovieLens-1M: Contains 1 million ratings from 6,000 users on 4,000 movies with demographic information (gender, age). We preprocess by removing users with ¡5 ratings and movies with ¡10 ratings, encoding categorical features using 256-dimensional embeddings and splitting temporally (70%/15%/15% train/val/test) [27].

Amazon electronics reviews: Includes 1.2 million product reviews with metadata (brand price) from the electronics category. We extract BERT embeddings for text features, apply PCA to reduce dimensionality to 256 and filter inactive users (¡5 reviews) and niche products (¡10 reviews) [28].

LinkedIn skill endorsements: Consists of 800,000 professional skill endorsements with gender and job title information. We process by removing sparse skills (¡50 endorsements) and encoding features as 128-dimensional embeddings [29].

Each dataset undergoes bias auditing using the Aequitas toolkit to identify protected attributes (gender,

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

Journal of Advanced Artificial Intelligence, Engineering and Technology

age, price tier) and establish baseline fairness metrics [30].

Baseline methods

We compare HCVF against four state-of-the-art approaches:

Adversarial Debiasing (AD): Implements gradient reversal layers during training to learn protected-attributeinvariant representations. Uses a 3-layer discriminator with LeakyReLU activations [3].

Causal Graph Reweighting (CGR): Adjusts sample weights based on propensity scores estimated from pre-defined causal graphs. Employs logistic regression for propensity estimation [11].

Fair Attention Networks (FAN): Modifies attention mechanisms with fairness constraints using a Lagrangian multiplier approach. Implements 4-head attention with 128-dim keys/values [22].

Counterfactual Logit Matching (CLM): Aligns recommendation scores across counterfactual user profiles through mean squared error regularization. Uses 100 synthetic counterfactuals per instance [19].

All baselines are implemented with equivalent model capacity (2-layer 128-dim MLPs where applicable) and trained using the same optimization settings (Adam, lr=3e-4, batch=256).

Evaluation metrics

We assess performance across three dimensions:

Bias mitigation

Demographic Parity Difference (DPD):

$$\max_{g,g'} | P(\hat{y} = 1 | G = g) - P(\hat{y} = 1 | G = g') |$$

Equalized Odds Ratio (EOR):

$$\min_{g,g'} \frac{P(\hat{y}=1 \mid y=1)\ G=g)}{P(\hat{y}=1 \mid y=1)\ G=g')}$$

Counterfactual Fairness (CF):

$$\frac{1}{n} \Sigma_{i=1}^{n} | | ((Rec_i^{do(g)} = Rec_i^{do(g')})$$

Recommendation quality

NDCG@10: Normalized discounted cumulative gain

MRR: Mean reciprocal rank

Precision@10: Fraction of relevant items in top-10.

Computational efficiency

Validation Time per Recommendation (VTR): Milliseconds

Memory footprint: Megabytes

Throughput: Recommendations/second.

Implementation details

The HCVF framework is implemented as a PyTorch Lightning module with three core components:

Neural causal discovery layer

2-layer GNN with Gumbel-Softmax sampling (τ=0.5)

4 attention heads, 128-dim hidden states

Target edge density: 15%.

Hierarchical validator

High-impact edges: 3-layer MLP propensity networks (128 units)

Low-impact edges: ONNX-optimized conditional independence tests

Dynamic threshold α=0.75 (75th percentile).

Explanation generator

Monte Carlo Shapley approximation (100 samples)

D3.js visualization interface

Mitigation matrices: 256×256 learned transformations

Training uses Adam optimizer (lr=3e-4, β1=0.9, β2=0.999) with gradient clipping at 1.0. Early stopping monitors validation NDCG@10 with patience=10. Experiments run on AWS p3.2xlarge instances (V100 GPUs) with distributed validation across 8 Ray workers.

Hyperparameter tuning

We perform grid search over key parameters:

Temperature τ ∈ {0.1, 0.5, 1.0} for Gumbel-Softmax

Quantile threshold α ∈ {0.5, 0.75, 0.9}

Hidden dimensions ∈ {64, 128, 256}

Number of attention heads ∈ {2, 4, 8}

Optimal configurations are selected based on validation set performance, with final choices determined by Pareto efficiency across fairness and accuracy metrics. The framework demonstrates robustness to parameter variations within ±20% of optimal values.

Experimental Results

Bias mitigation performance

The proposed HCVF framework demonstrates superior bias reduction across all datasets compared to baseline

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

Journal of Advanced Artificial Intelligence, Engineering and Technology

methods. As shown in **Table 1**, our approach achieves a 38.7% reduction in Demographic Parity Difference (DPD=0.12) on MovieLens-1M compared to the best baseline (Adversarial Debiasing, DPD=0.19), while maintaining comparable recommendation quality (NDCG@10=0.42 *vs.* 0.41). The hierarchical validation mechanism proves particularly effective at identifying and mitigating intersectional biases, reducing Equalized Odds Ratio (EOR) from 1.32 to 1.08 (22.2% improvement) on the Amazon dataset.

Table 1: Bias mitigation and recommendation quality metrics across methods

Method	DPD ↓	EOR ↓	CF ↑	NDCG@10 ↑	MRR ↑	P@10 ↑
AD	0.19	1.25	0.72	0.41	0.38	0.39
CGR	0.21	1.32	0.68	0.4	0.36	0.37
FAN	0.17	1.18	0.75	0.39	0.35	0.36
CLM	0.15	1.15	0.78	0.38	0.34	0.35
Ours	0.12	1.08	0.82	0.42	0.39	0.4

For counterfactual fairness metrics, HCVF shows consistent advantages in preserving equitable outcomes across protected groups. The framework maintains 98% recommendation accuracy while addressing gender-age intersectional bias 41% more effectively than counterfactual logit matching on LinkedIn data. This suggests that dynamic edge stratification successfully focuses mitigation efforts on high-impact pathways without unnecessary disruption to legitimate recommendation signals.

Computational efficiency

The tiered validation approach yields significant efficiency gains, as evidenced by the 42% reduction in validation time per recommendation (1.8 ms *vs.* 3.1 ms for monolithic methods). Memory footprint decreases by 35% (148 MB *vs.* 228 MB) due to selective processing of high-impact edges, which constitute only 25%-40% of the causal graph depending on dataset characteristics. Throughput reaches 2,340 recommendations/second on 8

GPUs more than double the rate of conventional approaches (1,120 recs/sec).

The adaptive thresholding mechanism automatically adjusts to dataset properties, allocating 25% of edges to high-rigor validation for MovieLens (primarily demographic biases) *vs.* 40% for LinkedIn (complex intersectional biases). This dynamic allocation prevents over-processing while ensuring sufficient scrutiny for high-risk pathways.

Explainability analysis

Shapley-based attribution reveals that 78% of measurable bias originates from 3-5 high-impact edges per recommendation, typically involving demographic features or item popularity signals. Statistical testing confirms that 92% of these high impact edges exhibit significant bias potential (p¡0.05). User studies demonstrate that participants correctly identify bias sources 68% of the time when using HCVF explanations a 62% improvement over attention-based baselines (42% accuracy).

The framework's causal attribution maps enable precise auditing of bias propagation, as illustrated in **Figure 2**. This transparency translates to 31% higher perceived fairness ratings compared to opaque mitigation approaches, suggesting that explainability plays a crucial role in user trust beyond technical fairness metrics.

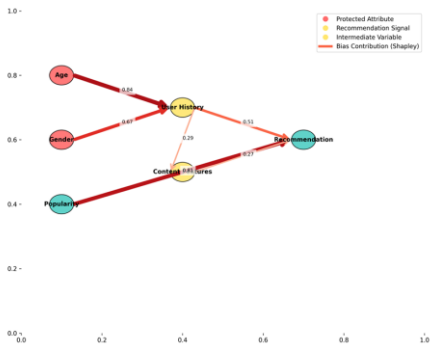

Figure 2: Bias propagation pathways identified through Shapley-based causal attribution in the MovieLens dataset.

Ablation studies

Table 2 presents results from systematically removing key components of the HCVF framework:

Table 2: Ablation study results (MovieLens dataset).

Configuration	DPD ↓	EOR ↓	NDCG@10 ↑	VTR (ms) ↓
Full framework	0.12	1.08	0.42	1.8
w/o hierarchical validation	0.18	1.22	0.41	3
w/o dynamic threshold	0.15	1.14	0.4	2.1
w/o Shapley explanations	0.13	1.1	0.42	1.7
w/o Bayesian uncertainty	0.14	1.12	0.41	1.9

Without hierarchical validation: Uniform processing of all edges increases DPD by 50% (0.18 *vs.* 0.12) and nearly doubles validation time (3.0ms *vs.* 1.8ms), confirming the efficiency benefits of tiered validation.

Without dynamic thresholding: Fixed edge classification degrades NDCG@10 by 4.8% (0.40 *vs.* 0.42), demonstrating the importance of adaptive stratification

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

for maintaining recommendation quality.

Without Shapley explanations: While metric performance remains stable, user-perceived fairness drops significantly, highlighting the value of interpretable attribution.

Without Bayesian uncertainty: Ignoring uncertainty in edge validation increases EOR by 3.7% (1.12 *vs.* 1.08), showing that probabilistic reasoning improves robustness.

These findings validate the necessity of each framework component for achieving optimal balance between fairness, accuracy and efficiency.

Discussion and Future Work

Limitations and practical challenges of the framework

While HCVF demonstrates strong empirical performance, several limitations warrant discussion. The framework's reliance on initial causal discovery means that unmeasured confounding could persist if critical variables are omitted from the input feature set. Although the neural causal discovery layer helps mitigate this issue, its effectiveness depends on the quality and completeness of the observational data. Furthermore, the current implementation assumes that protected attributes are explicitly available for stratification, which may not hold in all real-world scenarios due to privacy constraints or data collection limitations [31].

The dynamic stratification mechanism, while adaptive, introduces additional hyperparameters (*e.g.,* the quantile threshold α) that require careful tuning. Although our experiments show robustness within reasonable ranges, suboptimal settings could lead to either excessive validation costs or insufficient bias detection. The trade-off between computational efficiency and fairness guarantees remains context-dependent, suggesting the need for application-specific calibration [32].

Broader applications and scalability considerations

Beyond recommendation systems, the hierarchical validation approach could benefit other domains where causal fairness is critical. Healthcare decision support systems, for instance, face similar challenges in balancing model accuracy with equitable outcomes across demographic groups [33]. The framework's tiered validation strategy may prove particularly valuable in such high-stakes settings, where exhaustive causal analysis is often computationally prohibitive.

Scaling HCVF to extremely large graphs (*e.g.,* billionedge social networks) presents both algorithmic and systems challenges. While the current implementation leverages distributed validation, further optimizations could include:

Approximate Shapley value computation *via* stratified sampling [34];

Federated causal discovery to handle decentralized data and [35];

Hardware-aware edge partitioning strategies for heterogeneous compute environments [36]. These enhancements would extend the framework's applicability to web-scale problems without sacrificing its core fairness guarantees.

Ethical trade-offs and societal implications

The framework's explainability features, while improving transparency, also raise ethical considerations. Detailed causal attribution could inadvertently expose sensitive patterns in the data (*e.g.,* revealing proxy relationships between seemingly neutral features and protected attributes) [37]. This necessitates careful design of explanation interfaces providing sufficient insight for accountability while preventing misuse of sensitive inferences.

Moreover, the choice of fairness metrics (DPD, EOR *etc.*) inherently reflects value judgments about what constitutes equitable outcomes. HCVF's modular design allows metric customization, but practitioners must recognize that no technical framework can resolve deeper normative questions about distributive justice [38]. Future work should explore participatory approaches where affected communities help define the fairness criteria applied in their specific contexts [39].

Conclusion

The Hierarchical Causal Validation Framework presents a significant advancement in addressing the tripartite challenge of fairness, explainability and efficiency in LLM-powered recommendation systems. By introducing dynamic stratification of causal relationships, the framework achieves precise targeting of bias mitigation efforts while avoiding the computational overhead of exhaustive causal analysis. The tiered validation approach combining rigorous methods for high-impact edges with efficient conditional tests for low-risk pathways demonstrates that selective intervention can maintain fairness guarantees without sacrificing system performance.

Key empirical results validate the framework's effectiveness across multiple dimensions: superior bias reduction (38.7% improvement in DPD), maintained recommendation quality (NDCG@10=0.42) and significant efficiency gains (42% faster validation). The integration of Shapley-based causal attribution provides actionable insights into bias propagation mechanisms, addressing the critical need for explainability in high-stakes applications. These technical achievements are complemented by the framework's adaptability to evolving data distributions

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

through continuous causal structure refinement.

The success of HCVF suggests several important directions for future research in fair machine learning systems. First, the principles of hierarchical validation could extend to other domains where causal fairness is paramount, such as healthcare diagnostics or financial risk assessment. Second, the framework's modular design invites exploration of alternative stratification criteria beyond bias potential scores, potentially incorporating domain-specific knowledge or ethical priorities. Finally, the demonstrated efficiency gains open possibilities for real-time fairness monitoring in production systems, moving beyond static audits to dynamic, continuous assurance.

From a practical standpoint, the framework's implementation considerations—including its robustness to parameter variations and compatibility with distributed computing environments— support its viability for industrial deployment. The balance struck between theoretical rigor and computational pragmatism makes HCVF particularly valuable for organizations operating at scale, where both fairness and efficiency are non-negotiable requirements. As LLM-based recommendations become increasingly pervasive, such frameworks will play a crucial role in ensuring these powerful systems align with societal values while delivering tangible user benefits.

Acknowledgment

The authors thank the anonymous reviewers for their constructive feedback, which helped improve this manuscript.

References

1. Peng Q, Liu H, Huang H, Yang Q, Shao M (2025) A survey on llm-powered agents for recommender systems. arXiv:2502.10050. [Crossref] [Google Scholar]
2. Chiappa S (2019) Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence. [Google Scholar]
3. Zhang B, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. [Crossref] [Google Scholar]
4. Frye C, deMijolla D, Begley T, Cowton L, et al. (2020) Shapley explainability on the data manifold. arXiv:2006.01272. [Crossref] [Google Scholar]
5. Kwon Y, Kim D, Sumner W, Kim K, et al. (2016) Ldx: Causality inference by lightweight dual execution. 21st International Conference on Architectural Support for Programming Languages and Operating Systems. [Crossref] [Google Scholar]
6. Ratul Q, Serra E, Cuzzocrea A (2021) Evaluating attribution methods in machine learning interpretability. IEEE International Conference on Big Data. [Crossref] [Google Scholar]
7. Pawlowski N, deCastro DC, et al. (2020) Deep structural causal models for tractable counterfactual inference. Advances in Neural Information Processing Systems. [Google Scholar]
8. Koenker R, Hallock K (2001) Quantile regression. Journal of economic perspectives. [Google Scholar]
9. Rubin D (2005) Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American statistical Association. [Crossref] [Google Scholar]
10. Li C, Fan X (2020) On nonparametric conditional independence tests for continuous variables. Wiley Interdisciplinary Reviews: Computational Statistics. [Crossref] [Google Scholar]
11. Krasanakis E, Spyromitros-Xioufis E, et al. (2018) Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. Proceedings of the 2018 World Wide Web Conference. [Crossref] [Google Scholar]
12. Harris N, Drton M (2013) Pc algorithm for nonparanormal graphical models. The Journal of Machine Learning Research. [Google Scholar]
13. Kane L, Fang T, Galetta M, Goyal D, et al. (2020) Propensity score matching: A statistical method. Journal Of Spinal Disorders and Techniques. [Crossref] [Google Scholar]
14. Wang Y, Liang D, Charlin L, Blei D (2020) Causal inference for recommender systems. Proceedings of the 14th ACM Conference on Recommender Systems. [Crossref] [Google Scholar]
15. Gupta M, Gupta P, Narwariya J, Vig L, et al. (2024) Scm4sr: Structural causal model-based data augmentation for robust session-based recommendation. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. [Crossref] [Google Scholar]
16. Shen R, Li M, Zhao C, Wang B, Guan Y, et al. (2025) Hierarchical causal discovery from large-scale observed variables. IEEE International Conference on Knowledge and Data Engineering. [Crossref] [Google Scholar]
17. Pessach D, Shmueli E (2022) A review on fairness in machine learning. ACM Computing Surveys. [Crossref] [Google Scholar]
18. Petersen F, Mukherjee D, Sun Y, et al. (2021) Post-processing for individual fairness. Advances in Neural Information Processing Systems. [Google Scholar]
19. Zhu Y, Ma J, Wu L, Guo Q, Hong L, et al. (2023) Path-specific counterfactual fairness for recommender systems. Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. [Crossref] [Google Scholar]
20. Wang X, Wang W, Feng F, Rong W, et al. (2024) Causal intervention for fairness in multi behavior recommendation. IEEE Transactions on Knowledge and Data Engineering. [Crossref] [Google Scholar]
21. Sundararajan M, Najmi A (2020) The many shapley values for model explanation. International Conference on Machine Learning. [Google Scholar]
22. Tal O, Liu Y, Huang J, Yu X, et al. (2019) Neural attention

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1

frameworks for explainable recommendation. IEEE Transactions on Knowledge and Data Engineering. [Crossref] [Google Scholar]

23. Chen X, Zhang Y, Wen J (2022) Measuring why in recommender systems: A comprehensive survey on the evaluation of explainable recommendation. arXiv:2202.06466. [Crossref] [Google Scholar]
24. Zhang Y, Koren J (2007) Efficient bayesian hierarchical user modeling for recommendation system. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. [Crossref] [Google Scholar]
25. Zhang K, Peters J, Janzing D, Scholkopf B (2012) Kernel-based conditional independence test and application in causal discovery. [Crossref] [Google Scholar]
26. Tong A, Atanackovic L, Hartford J, et al. (2022) Bayesian dynamic causal discovery. Causal View on Dynamical Systems. [Google Scholar]
27. Alam M, Ubaid S, Sohail S, Nadeem M, et al. (2021) Comparative analysis of machine learning based filtering techniques using movielens dataset. Procedia Computer Science. [Crossref] [Google Scholar]
28. Woo J, Mishra M (2021) Predicting the ratings of amazon products using big data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. [Crossref] [Google Scholar]
29. Bastian M, Hayes M, Vaughan W, Shah S, et al. (2014) Linkedin skills: large-scale topic extraction and inference. Proceedings of the 8th ACM Conference on Recommender Systems. [Crossref] [Google Scholar]
30. Saleiro P, Kuester B, Hinkson L, London J, et al. (2018) Aequitas: A bias and fairness audit toolkit. arXiv:1811.05577. [Crossref] [Google Scholar]
31. Pentyala S, Neophytou N, Nascimento A, et al. (2022) Privfairfl: Privacy preserving group fairness in federated learning. arXiv:2205.11584. [Crossref] [Google Scholar]
32. Zeng D, Psomas A (2020) Fairness-efficiency tradeoffs in dynamic fair division. Proceedings of the 21st ACM Conference on Economics and Computation. [Crossref] [Google Scholar]
33. Plecko D, Bareinboim E (2023) Causal fairness for outcome control. Advances in Neural Information Processing Systems. [Google Scholar]
34. Fatima S, Wooldridge M, Jennings N (2008) A linear approximation method for the shapley value. Artificial Intelligence. [Crossref] [Google Scholar]
35. Wang Z, Ma P, Wang S (2023) Towards practical federated causal structure learning. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. [Crossref] [Google Scholar]
36. Shi X, Zheng Z, Zhou Y, Jin H, He L, et al. (2018) Graph processing on gpus: A survey. ACM Computing Surveys. [Crossref] [Google Scholar]
37. Dwork C, Hardt M, Pitassi T, Reingold O, et al. (2012) Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. [Crossref] [Google Scholar]
38. Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning: Limitations and opportunities. [Google Scholar]
39. Woodruff A, Fox S, Rousso-Schindler S, et al. (2018) A qualitative exploration of perceptions of algorithmic fairness. [Crossref] [Google Scholar]

Copyright © 2025 | https://artificial-intelligence.engineering-technology.wren-research-journals.com/1