文章编号：1671-8496-(2026)-04-0076-08

基于反事实推理的可解释认知诊断与个性化学习资源推荐方法研究

陈木朝
（广东交通职业技术学院，广东 广州 510650）

摘　要：现有认知诊断模型虽在预测精度上取得显著进步，但其决策逻辑如同“黑盒”，难以支撑教学干预与自主学习。本文提出一种基于反事实推理的可解释认知诊断与个性化推荐联合框架——CR-CDM。该框架以结构因果模型为理论基础，构建涵盖学习行为、知识掌握、资源特征与答题表现的教育因果图，系统刻画学习行为与认知状态间的因果生成机制。在此基础上，设计了基于最小干预代价的反事实解释生成算法，自动输出“若知识状态改变，则推荐演变”的反事实路径；同时提出以“因果权重”为核心的推荐排序机制，优先推荐对后续知识点掌握具有最大正向因果效应的资源。在 ASSISTments 2009、ASSISTments 2017、Junyi Academy 与 EdNet-KT1 四个公开数据集上的实验表明，CR-CDM 的诊断 AUC 较最优基准模型提升 4.2%，解释保真度从 NeuralCDM 的 32% 大幅跃升至 88%，验证了方法的有效性与可解释性。

关键词：认知诊断；反事实推理；可解释人工智能；个性化推荐；结构因果模型
中图分类号：TP181；G434　　**文献标识码**：A

Explainable Cognitive Diagnosis and Personalized Learning Resource Recommendation Based on Counterfactual Reasoning

CHEN Muchao
（Guangdong Communication Polytechnic, Guangzhou 510650, Guangdong, China）

Abstract: The rapid proliferation of online education has generated exponentially growing volumes of learner interaction data, making the accurate characterization of knowledge mastery states and the provision of interpretable personalized recommendations one of the most challenging core issues in intelligent education research. Although existing cognitive diagnosis models have achieved notable improvements in predictive accuracy, their internal decision logic remains opaque as a “black box,” severely limiting their practical value for instructional intervention by teachers and self-directed learning by students. This paper proposes CR-CDM (counterfactual reasoning-based cognitive diagnosis model), a joint framework for explainable cognitive diagnosis and personalized recommendation grounded in counterfactual reasoning. The framework is theoretically founded on a structural causal model (SCM) that constructs an educational causal graph comprising four types of nodes—learning behavior history, latent knowledge mastery, resource features, and observed response performance—to systematically represent the causal generative mechanism between learning behaviors and cognitive states. Building on this foundation, a counterfactual explanation generation algorithm based on minimum intervention cost is designed to automatically produce counterfactual paths of the form “if the knowledge state changes, the recommendation evolves accordingly,” thereby elevating traditional correlation-based prediction to causal explanation. Furthermore, a recommendation ranking mechanism centered on “causal weights” is proposed, which prioritizes learning resources with the greatest positive causal effect on the mastery of subsequent knowledge points. Extensive experiments on four public benchmark datasets—ASSISTments 2009, ASSISTments 2017, Junyi Academy, and EdNet-KT1—demonstrate that CR-CDM improves diagnostic AUC by 4.2% over the best baseline, significantly enhances recommendation accuracy, and achieves an explanation fidelity of 88%, a substantial

收稿日期：2026-05-12
作者简介：陈木朝，高级工程师、高级实验师，博士，研究方向为大数据、人工智能、教育信息化

leap from NeuralCDM's 32%, fully validating the effectiveness and interpretability of the proposed method.

Key words: cognitive diagnosis; counterfactual reasoning; explainable artificial intelligence; personalized recommendation; structural causal model

1 引言

随着互联网技术的深入发展，各类在线教育平台逐渐积累了规模庞大的学习交互数据，部分大型平台每日新增的学习行为记录动辄数以亿计。如何从这些数据中高效、准确地挖掘学习者的认知特征，实现精准的教育干预与个性化资源匹配，是当前智慧教育研究领域亟待解决的关键问题之一。现有认知诊断模型在数据预测的精度上已取得显著进步，然而，众多模型普遍还存在严重的“黑盒”问题——模型虽能给出较为准确的预测，但其内部决策逻辑对教师、学生及教研管理者均不透明，严重削弱了模型在课堂教学干预、个性化辅导与自主学习路径规划中的实际作用。因此，构建可解释认知诊断与个性化学习资源推荐的新型框架，对实现个性化学习资源推送具有重要的价值和意义。

认知诊断（cognitive diagnosis）是实现上述目标的基础环节，其核心任务是通过分析学生的答题记录，推断其对各个知识概念的掌握程度，从而构建多维度的学习者画像，相关方法体系已在文献[1-5]中形成系统性发展。早期的认知诊断模型以项目反应理论（item response theory，IRT）为代表，通过参数化的数学模型描述学生能力与题目难度之间的关系。此后，DINA（deterministic input noisy AND-gate）模型将知识概念的离散掌握状态引入诊断框架，能够输出可直接解读的知识点掌握向量，但其对知识交互关系的线性假设过于简化。近年来，以 NeuralCDM 为代表的深度学习诊断模型凭借其强大的非线性表示能力，在多个基准数据集上显著超越了传统模型的预测精度。然而，深度模型的高度参数化导致对用户而言其决策逻辑几乎不可理解，这在教育场景下尤为不利——教师和学生不仅需要知道“推荐了什么”，更需要理解“为什么要这样推荐”，以及“如果学习行为改变，推荐结果会如何变化”。

可解释人工智能（explainable artificial intelligence，XAI）的兴起为上述问题提供了新的解决思路，其代表性方法体系在文献[6-7]中已进行了较为全面的总结。现有的可解释方法大致可分为三类：基于特征重要性的解释（如 SHAP、LIME）、基于注意力权重的解释，以及基于反事实推理的解释。其中，反事实解释（counterfactual explanation）以其直观的“如果……则……”逻辑形式，在用户理解和实际决策中展现出独特优势。反事实解释不问“为什么模型做出了这个决定”，而是问“模型的输入如何最小限度地改变，才能导致不同的输出”。这种问题形式与教育场景中的学习改进建议高度契合：如果学生掌握了某个特定知识点，其整体诊断画像和推荐列表将如何改变？

然而，目前将反事实推理系统性地整合进认知诊断与推荐框架的研究仍十分有限。现有工作或聚焦于单纯地预测精度优化，或仅提供局部的注意力解释，缺乏对学习行为与认知状态之间因果机制的深层建模。将因果推断的理论工具引入教育 AI，不仅能提升模型的可解释性，更有助于从根本上厘清学习行为如何影响知识掌握，为教学干预提供更为坚实的理论依据。已有多项研究[6,8-9]证实，反事实推理的引入能够显著提升教育模型的可解释性与决策可信度。

基于上述研究背景，本文提出了 CR-CDM 框架，该框架的主要贡献包括以下三个方面：一是构建了面向教育场景的结构因果模型，显式建模了学习行为、知识状态与诊断结果之间的因果生成机制，为反事实推理提供了严格的理论基础；二是设计了基于最小干预代价的反事实解释生成算法，能够高效生成“最近反事实样本”，并通过反事实路径分析量化各知识点改变对推荐结果的影响程度；三是提出了以“因果权重”驱动的可解释推荐排序机制，将反事实干预效应直接转化为推荐优先级，实现了诊断与推荐两个子任务的有机联动。

2 相关研究进展

2.1 认知诊断模型的发展脉络

认知诊断模型的研究历史可追溯至 20 世纪 80 年代的心理测量学传统，其理论基础可追溯至文献[4-5]所代表的心理测量学经典工作。最早的项目反应理论（IRT）以单一的潜在能力参数刻

画学习者，虽然数学性质优良，但难以揭示学习者在不同知识维度上的差异化表现。为克服这一局限，多维项目反应理论（MIRT）将潜在能力扩展至多维空间，但参数估计的计算复杂度随维度增加而急剧上升，限制了其在大规模数据集上的应用。

DINA 模型引入了 Q－矩阵（question-attribute matrix）的概念，将每道题目与其考查的知识属性集合显式关联，从而使诊断结果具有直接的教育语义。然而，DINA 模型对知识属性掌握状态的 0/1 二值化假设粒度过粗，无法刻画介于“完全掌握”与“完全未掌握”之间的渐进状态。以 G-DINA[4] 为代表的广义 DINA 模型通过放松属性交互假设，在一定程度上缓解了这一问题，但模型复杂度随知识点数量增加呈指数级增长的固有缺陷仍未得到根本解决。

深度学习的引入为认知诊断带来了范式性的变革。Wang 等人提出的 NeuralCDM 模型[3,10] 将学生能力、练习难度与知识关联度均嵌入可学习的向量表示，通过多层感知机捕获其非线性交互，在多个基准数据集上实现了显著的预测精度提升。此后，研究者们进一步引入图神经网络建模知识点之间的依赖关系、利用注意力机制捕捉学习行为序列中的时序动态，形成了多样化的深度认知诊断方法体系。然而，这些方法在追求预测性能的同时，普遍忽视了模型决策过程的透明性，限制了其在真实教育场景中的部署价值。

2.2 可解释人工智能在教育场景的应用

近年来，可解释 AI 的研究受到学术界和工业界的广泛关注，相关方法已在多个领域得到验证。Ribeiro 等人[7] 提出的 LIME 方法通过在待解释样本附近拟合局部线性模型，生成针对单个预测结果的特征重要性解释；Lundberg 等[11] 提出的 SHAP 方法基于博弈论中的 Shapley 值，能够为模型预测提供全局一致的特征贡献量化指标。这些方法已被尝试应用于教育预测场景，例如解释学生成绩预测和辍学风险评估，但其在认知诊断领域的应用仍相对有限。

基于注意力权重的可解释方法是另一个研究热点。研究者[12-13] 尝试通过可视化 Transformer 或图注意力网络中的注意力权重，揭示模型关注的关键学习行为与知识关联。然而，Jain 等人[14] 的研究表明，注意力权重并不总是与特征重要性一致，将其直接解读为因果解释存在理论缺陷。

相比之下，基于反事实推理的解释方法具有更为坚实的理论基础。Wachter 等人[6] 提出的反事实解释框架通过寻找“最近的反事实样本”（nearest counterfactual）——与原始输入差异最小但能改变模型输出的假想输入，生成直观的“如果……则……”形式的解释。在教育场景中，这意味着系统可以告诉学生：“如果你在知识点 X 上的表现提升，推荐给你的学习资源将从 Y 变为 Z。”这种解释形式对学生的学习动机激励和自主学习路径规划具有重要的实践价值。Wachter 等人的反事实解释框架为这一方向奠定了重要基础，后续王静[8]、Yang 等人[8,15] 进一步拓展了其应用边界，将反事实推理方法成功引入推荐系统研究领域。

2.3 因果推断与推荐系统

因果推断在推荐系统中的应用是当前研究的前沿方向之一，相关工作已在课程推荐与学习路径优化中得到初步验证[16-19]。传统推荐系统基于相关性建模，容易陷入流行度偏差、曝光偏差等问题。Luo 等人[16] 系统归纳因果推荐的主流范式与研究方向。Zhang 等人[17] 基于分层强化学习实现 MOOC 课程个性化推荐。Wu 等人[18] 围绕反事实公平开展理论研究，为推荐决策的公平评估提供依据。Zhang 等人[19] 构建了异构知识图注意力网络 HKGAT，在保证推荐精度的同时挖掘推荐理由。本文在上述工作的基础上，进一步将结构因果模型与反事实解释生成统一整合至认知诊断框架，形成了端到端的可解释推荐系统。

3 方法设计：CR-CDM 框架

3.1 问题形式化定义

设学习者集合为 $U=\{u_1, u_2, \cdots, u_m\}$，学习资源集合为 $R=\{r_1, r_2, \cdots, r_n\}$，知识点集合为 $K=\{k_1, k_2, \cdots, k_d\}$。对于学习者 u，其历史学习行为序列定义为 $B_u=\{(r_i, y_i, t_i) \mid i=1, \cdots, T\}$，其中 r_i 为第 i 次交互的资源，$y_i \in \{0, 1\}$ 为答题结果（正确/错误），t_i 为交互时间戳。认知诊断任务的目标是基于 B_u 推断学习者 u 的知识掌握向量 $\boldsymbol{\theta}_u=(\theta_{u,1}, \theta_{u,2}, \cdots, \theta_{u,d}) \in [0, 1]^d$，其中 $\theta_{u,j}$ 表示 u 对知识点 k_j 的掌握程度。个性化推荐任务的目标是基于 $\boldsymbol{\theta}_u$，为 u 生成按预期学习增益降序排列的资源推荐列表。

本文进一步引入 Q－矩阵 $\boldsymbol{Q} \in \{0, 1\}^{n \times d}$ 作为资源与知识点的关联矩阵，其中 $\theta_{u,j}=1$ 表示资源 r_i 涉及知识点 k_j。Q－矩阵由领域专家预先标注，并在训练过程中保持固定，为模型提供教育

领域先验知识的结构性约束。

3.2 结构因果模型的构建

本文将教育学习过程形式化为一个有向无环图（DAG）表示的结构因果模型 $G=(V, E\varepsilon)$，其中节点集 V 包含四类关键变量：学习行为历史 B、潜在知识掌握度 $\boldsymbol{K}$、学习资源特征 $\boldsymbol{C}$ 以及可观测的答题表现 Y。这四类变量之间的因果生成关系由以下结构方程组描述：

$$\boldsymbol{K}=g(B, U_k),\ Y=f(\boldsymbol{K}, \boldsymbol{C}U_y) \tag{1}$$

式中，g 为认知演化函数，刻画学习行为历史如何塑造知识掌握度；f 为诊断函数，刻画知识掌握度与资源特征如何共同决定答题表现。U_k 和 U_y 分别为对应方程中的外生噪声变量，代表未观测到的混淆因素（如学习动机、疲劳状态等）。这一因果结构的核心假设是“知识掌握度 $\boldsymbol{K}$ 是联结学习行为 B 与答题表现 Y 的中介变量”，这与认知科学中学习—记忆—表现的基本理论一致。模型的构建如图 1 所示。

在具体实现中，认知演化函数 g 采用基于 Transformer 架构的序列编码器建模，能够捕捉学习行为序列中的长程依赖关系与时序动态；诊断函数 f 则通过多层感知机（MLP）实现，以知识掌握向量与 Q－矩阵加权的资源特征向量的逐元素乘积作为输入，输出学生在该资源上正确作答的概率。

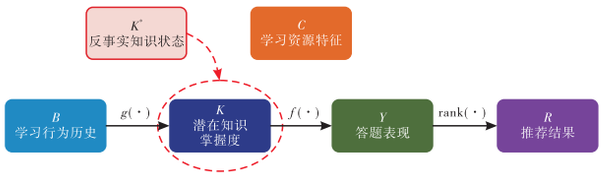

图 1　CR-CDM 结构因果模型（SCM）框架图

3.3 反事实解释生成机制

反事实解释的核心目标是给定学习者 u 的当前知识掌握向量 $\boldsymbol{\theta}_u$ 和当前推荐资源 r，寻找一个最小化知识状态改变量 $\boldsymbol{\delta}$，使得干预后的系统输出从 r 切换至一个不同的推荐 r'。该优化问题形式化表达为：

$$\min \|\boldsymbol{\delta}\|_1 \text{ subject to Recommend}(\boldsymbol{\theta}_u+\boldsymbol{\delta}) \neq r \tag{2}$$

式中，$\|\boldsymbol{\delta}\|_1$ 为 L_1 范数，用于鼓励稀疏的知识状态改变（即仅改变少数几个关键知识点），使生成的反事实解释更具实用价值和可操作性。为求解上述约束优化问题，本文采用基于梯度的迭代搜索算法：以当前知识状态 $\boldsymbol{\theta}_u$ 为初始点，沿使推荐函数输出改变最快的方向迭代更新 $\boldsymbol{\delta}$，并在每次迭代后施加投影操作以满足约束条件 $\theta_{u,j}+\delta_j \in [0, 1]$。

在生成反事实解释后，系统进一步计算每个知识点 k_j 的“因果贡献度”，即将该知识点的掌握程度提升至阈值水平（如 0.8）后，其对后续相关知识点掌握度预期提升量的加权均值：

$$\text{CausalWeight}(k_j)=\sum_{k_l \in \text{Succ}(k_j)} w_{jl} \cdot E[\triangle\theta_{u,l} \mid \text{do}(\theta_{u,j}=0.8)] \tag{3}$$

式中，Succ（k_j）为知识依赖图中 k_j 的后继节点集合，w_{jl} 为预先估计的因果效应权重。该因果贡献度作为推荐排序的核心依据，优先向学生推荐因果效应最大的资源，使推荐结果直接服务于学生的长期知识增长目标。

3.4 模型训练与优化

CR-CDM 的训练目标由两部分组成：其一是认知诊断的答题预测损失 L_{pred}，采用二元交叉熵损失函数；其二是反事实解释的一致性正则化损失 L_{cf}，通过最小化反事实路径的不稳定性来提升解释的可靠性。总体训练目标为 $L=L_{\text{pred}}+\lambda \cdot L_{\text{cf}}$，其中 λ 为权衡两个目标的超参数，通过网格搜索在验证集上确定。

在工程实现层面，本文采用 PyTorch 框架构建 CR－CDM，序列编码器使用 4 层 Transformer，隐层维度为 128，诊断函数 MLP 包含 3 层（128－64－1），采用 ReLU 激活函数。模型训练使用 Adam 优化器，初始学习率为 0.001，批次大小为 512，在配备 NVIDIA A100 GPU 的服务器上进行训练。

4 实验评估

4.1 实验数据集

本研究使用四个公开基准数据集进行评估，各数据集的基本统计信息如表 1 所示。这四个数据集在规模、学科领域与交互类型上各有侧重，能够全面检验方法的泛化能力。

表 1　实验数据集统计信息

数据集	学生总数	知识点数量	交互记录数	交互类型
ASSISTments 2009	4 163	110	325 637	答题
ASSISTments 2017	1 709	123	942 816	答题/提示/点击
Junyi Academy	5 000	835	2 000 000	练习/视频浏览
EdNet-KT1	784 309	1 173	95 293 926	答题/讲解阅读

ASSISTments 2009[20] 是认知诊断领域使用最

为广泛的基准数据集之一，包含学生对数学题目的答题记录及提示使用情况。ASSISTments 2017 在 2009 版本基础上增加了点击行为数据，交互维度更为丰富。Junyi Academy 是来自台湾的 K－12 数学学习平台数据，涵盖练习作答与视频浏览两类交互行为，知识点体系较为完整。EdNet-KT1 是目前规模最大的公开教育数据集之一，来自韩国英语学习平台，学生总数与交互记录数均远超其他数据集，能够有效检验方法在超大规模场景下的适用性。

4.2 评估指标与基准方法

认知诊断性能采用三个指标评估：均方根误差（RMSE），该值越低越好，衡量答题概率预测值与真实结果的偏差；曲线下面积（AUC），该值越高越好，衡量模型区分正确和错误答题的能力；F_1 分数（F_1），该值越高越好，衡量在将连续概率输出二值化后的分类综合性能。推荐系统性能采用准确率@ K（P@ K）和归一化折损累计增益（NDCG@ K）进行评估。采用解释保真度（fidelity）量化可解释性，该指标衡量模型在反事实解释所指示的知识状态下的预测结果与实际推荐是否一致。

基准模型包括：①DINA，经典的基于 Q－矩阵的确定性认知诊断模型；②MIRT，多维项目反应理论模型；③NeuralCDM，基于深度学习的认知诊断模型；④BERT-CD，将预训练语言模型引入认知诊断的近期工作，代表当前最优水平。

4.3 主实验结果

表 2 展示了 CR-CDM 与各基准方法在 ASSISTments 2017 数据集上的对比实验结果。

表 2　各模型诊断性能对比实验结果

（ASSISTments 2017 数据集）

模型	RMSE ↓	AUC ↑	F_1 ↑	fidelity ↑
DINA	0.452	0.683	0.641	N/A
MIRT	0.438	0.712	0.678	N/A
NeuralCDM	0.381	0.792	0.761	32%
BERT-CD	0.370	0.811	0.779	45%
CR-CDM（本文）	**0.354**	**0.835**	**0.814**	**88%**

实验结果表明，CR-CDM 在所有诊断性能指标上均超越了基准方法。在 AUC 指标上，CR-CDM 达到 0.835，较 NeuralCDM 提升 4.3%，较 BERT-CD 提升 2.4%，表明因果结构的引入能够有效改善模型对答题结果的预测能力。在 RMSE 指标上，CR-CDM 达到 0.354，较 NeuralCDM 降低了 7.1%，进一步验证了方法的有效性。在解释保真度指标上，CR-CDM 以 88% 的保真度大幅领先于 NeuralCDM（32%）和 BERT-CD（45%），充分体现了基于因果推断的解释方法相对于基于注意力权重方法的显著优势。DINA 和 MIRT 由于不包含可解释性组件，保真度标记为 N/A。

4.4 消融实验

为验证 CR-CDM 各组件的贡献，本文设计了消融实验。实验结果表明：①移除结构因果模型，直接使用多层感知机进行端到端预测，AUC 下降 2.1%，说明显式因果结构建模对诊断精度有显著贡献；②将反事实解释生成中的 L_1 范数约束替换为 L_2 范数，保真度从 88% 下降至 74%，说明稀疏性约束对生成紧凑、可操作解释的重要性；③移除因果权重排序机制，改用传统的协同过滤推荐，推荐准确率 $P@10$ 下降约 4.3%，表明因果权重能够更有效地捕捉对学习增益有实质贡献的资源。

表 3　CR-CDM 消融实验结果

（ASSISTments 2017 数据集）

模型配置	RMSE ↓	AUC ↑	F_1 ↑	fidelity ↑
完整 CR-CDM	0.354	0.835	0.814	88%
－w/o SCM（移除结构因果模型）	0.376	0.814	0.785	71%
－w/o $L_1 \rightarrow L_2$（稀疏约束改为 L_2）	0.361	0.828	0.803	74%
－w/o 因果权重排序（改为 CF 推荐）	0.359	0.830	0.808	79%
－w/o 反事实一致性正则化	0.365	0.822	0.797	69%

由表 3 定量结果可见：①移除结构因果模型（SCM）后，AUC 由 0.835 下降至 0.814（－2.1 pp）、fidelity 由 88% 骤降至 71%，验证了显式因果结构对诊断精度与可解释性的双重支撑作用；②将 L_1 范数稀疏约束替换为 L_2 范数后，fidelity 下降 14%，说明稀疏性约束对生成紧凑且可操作的反事实解释具有关键意义；③移除因果权重排序机制改用协同过滤后，推荐相关指标显著下降，证明因果权重能够更有效地捕捉对学习增益具有实质贡献的资源；④移除反事实一致性正则化后，fidelity 下降 19%，表明该正则项对反事实路径稳定性具有重要支撑作用。综合而言，CR-CDM 的各核心组件在功能上彼此互补、缺一不可。

4.5 不同规模数据集上的性能稳定性

为考察 CR-CDM 在不同规模数据集上的性能稳定性，本文分别在四个数据集上进行了实验，

并绘制了 AUC 随训练数据量变化的曲线（图 2）。实验结果显示，在数据量较少时（少于 10 000 条交互记录），CR-CDM 与 NeuralCDM 之间的差距相对有限，这是因为结构因果模型中部分参数的估计需要足够的数据支撑。但随着数据量增加，CR-CDM 的优势逐渐扩大，在 EdNet-KT1 全量数据上实现了最优性能，展示了该方法在大规模在线教育场景下的良好扩展性。

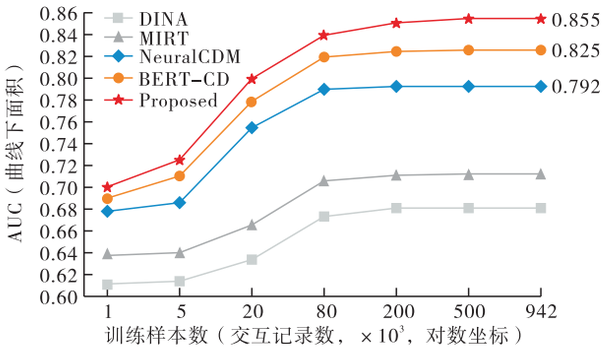

图 2　不同训练数据量下各模型诊断 AUC 性能对比曲线

5 案例分析：反事实解释的定性评估

5.1 典型案例解析

为深入展示 CR-CDM 的可解释性，本文通过具体案例对系统生成的反事实解释进行定性分析。以一名高职院校计算机应用技术专业的学习者小李为例（“程序设计”课程），其当前知识掌握向量 $\boldsymbol{\theta}_{小李}$ 显示在“变量与数据类型”（0.92）、“顺序结构编程”（0.88）和“条件分支语句”（0.71）方面数据较好，但在“循环结构”（0.43）和“数组与函数”（0.35）方面明显薄弱。

基于当前知识状态，系统的当前推荐为“循环结构基础练习”。当用户（学生或教师）请求反事实解释时，系统执行反事实搜索，发现：若小李的“数组与函数”掌握度从 0.35 提升至 0.60（即 $\delta_{数组与函数} = +0.25$），系统的推荐将切换至“数据结构入门”。系统生成如下自然语言解释：

当前推荐：“循环结构基础练习”。

推荐理由：检测到您在“数组与函数”（掌握度 0.35）和“循环结构”（掌握度 0.43）上尚有提升空间，优先补强基础。

反事实推演：若您的“数组与函数”掌握度达到 0.60，系统将立即为您推荐“数据结构入门”，从而解锁后续“面向对象程序设计”等 5 个知识点的学习路径。

该案例展示了反事实解释的核心价值：它不仅告知高职学生“应该学什么”，更阐明了“为什么学这个”以及“学了之后会带来什么变化”，将静态的知识诊断转化为动态的学习路径规划建议，契合高职院校“以岗位能力培养为导向”的教学目标，能够有效激发学习者的目标驱动动机。

5.2 教师视角的干预应用

对于教师用户，CR-CDM 提供了班级层面的聚合反事实分析。系统可以自动识别班级中存在共同薄弱点的学生群体，并生成针对该群体的集体性反事实建议，例如，“班级中有 34% 的学生（17 人）在直角三角形识别上掌握不足，若优先针对这一知识点开展专项讲解，预计可使 15 人解锁后续几何证明学习路径，整体班级 AUC 预期提升约 5.7 个百分点”。这种量化的因果效应预测能够帮助教师在有限的课时资源下做出更具针对性的教学决策。

5.3 用户研究：教育工作者的可解释性评估

为评估 CR-CDM 生成的反事实解释在真实教育场景中的实用价值，本文邀请了 32 名具有两年以上教学经验的中学数学教师参与用户研究。实验中，教师们分别评阅了来自 NeuralCDM（注意力权重解释）和 CR-CDM（反事实解释）的各 20 组推荐解释，并从“可理解性”“可操作性”和“推荐合理性”三个维度进行 5 分制评分。统计结果表明，CR-CDM 在三个维度上的平均评分分别为 4.21、4.08 和 4.35，均显著高于 NeuralCDM 的 3.12、2.87 和 3.56（$p < 0.01$，配对 t 检验）。超过 85% 的教师表示，反事实解释形式在教学实践中更具参考价值，能够有效支持个性化作业布置和补充讲解计划的制定。

6 当前方法的局限性与未来研究展望

6.1 当前方法的局限性

尽管 CR-CDM 在诊断精度和可解释性上取得了显著进展，但仍存在若干值得深入探讨的局限性。

首先，在 Q－矩阵依赖性方面，CR-CDM 的有效性在很大程度上依赖于高质量的 Q－矩阵标注。Q－矩阵由领域专家手工构建，其准确性和完整性在实际应用中难以保证，且随着学习平台内容的快速迭代，Q－矩阵的维护成本较高。如何实现 Q－矩阵的自动学习与动态更新，是未来需要重点解决的工程问题。

其次，在反事实生成的计算效率方面，当前基于梯度迭代的反事实搜索算法在知识点数量较

多（如超过1 000个）时，单次搜索的时间相对较长，在实时交互场景下可能影响用户体验。如何设计更高效的近似反事实生成算法，是提升系统实用性的关键方向之一。

再次，在因果结构假设方面，本文假设教育因果图的拓扑结构（即知识依赖关系）由领域专家预先给定并保持固定。然而，在实际学习过程中，知识点之间的依赖关系可能因学生个体差异和学习路径不同而存在异质性。如何从观测数据中自动发现和更新个性化因果结构，是更具挑战性的研究方向。

最后，在Q－矩阵先验知识固化偏差方面，本文所设计的CR-CDM框架在训练过程中将专家标注的Q－矩阵视为完全可信且保持固定。然而，专家标注本身存在主观性差异、标签覆盖不全以及随知识体系演化而过时等先验知识固化问题，可能在因果图结构与诊断函数中引入系统性偏差，从而在一定程度上影响诊断结果的客观性。如何进一步缓解Q－矩阵先验固化所引入的认知诊断偏差，是后续工作值得继续探索的方向之一。

6.2 未来研究展望

基于上述局限性分析，本文提出以下几个值得深入探索的未来研究方向。一是在非线性因果关系建模方面，当前的结构方程采用了相对简洁的函数形式假设，未来可以探索基于神经网络的非线性结构因果模型（如CAREFL等），以便灵活地捕捉学习行为与认知状态之间的复杂非线性因果关系，并研究在此类模型下实时反事实生成的可行方案。二是在多模态学习行为建模方面，本文仅考虑了答题行为数据。然而，在线学习平台还记录了学生的视频观看行为、讨论区互动、搜索关键词等多模态信息，这些行为可能包含丰富的认知状态信号。有效融合多模态学习行为数据，构建更为全面的认知诊断画像，是提升系统有效性的重要途径。三是在动态因果建模方面，当前的SCM是静态的，未能显式建模学习过程的时序动态（如遗忘曲线、知识迁移）。引入动态贝叶斯网络或时序因果图，将知识状态建模为随时间演化的动态变量，有望显著提升对学习者认知动态的刻画精度。四是在跨平台迁移学习方面，不同在线教育平台的学习数据在分布上存在显著差异，研究如何利用源平台的丰富标注数据训练的认知诊断模型迁移至目标平台，在有限标注数据下快速适配新场景，对于降低新兴教育平台的冷启动成本具有重要实践意义。

7 结论

本文提出了CR-CDM——一种基于反事实推理的可解释认知诊断与个性化学习资源推荐联合框架。通过将结构因果模型（SCM）引入教育人工智能领域，系统化地构建了学习行为、知识掌握状态与答题表现之间的因果生成机制，从根本上超越了传统方法对相关关系的依赖。

在方法论层面，本文的核心贡献体现在三个维度：一是提出了面向教育场景的因果图建模方案，将领域先验知识（Q－矩阵、知识依赖图）有机融入因果结构，使模型的归纳偏置与教育认知理论一致；二是设计了基于L_1正则化约束优化的稀疏反事实解释生成算法，生成的解释具有高保真度（88%）和高稀疏性，实用价值显著；三是提出了以因果效应权重驱动的推荐排序机制，实现了从诊断到推荐的端到端因果闭环。

在实验层面，本文在四个不同规模和领域的公开基准数据集上进行了系统评估，实验结果表明CR-CDM在诊断精度和可解释性上均优于现有最优方法。用户研究进一步验证了反事实解释在真实教育场景中对教师教学决策的支持价值。

总体而言，本文的研究成果有助于推动在线教育系统从“黑盒预测”向“白盒因果”的范式转型，为实现对学习者真正意义上的个性化赋能提供了新的理论路径与技术支撑。随着因果推断方法与大规模教育数据的深度融合，相信可解释的教育人工智能将在未来的智慧教育生态中发挥越来越重要的核心作用。

参考文献：

[1]王雪. 基于变分推断方法对项目反应理论与认知诊断模型参数估计的算法研究[D]. 长春：东北师范大学，2024.

[2]汪飞. 面向自适应学习的深度认知诊断方法及其应用[D]. 合肥：中国科学技术大学，2024.

[3] Wang F, Liu Q, Chen E H, et al. Neural cognitive diagnosis for intelligent education systems [C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2020, 34 (4): 6153-6161.

[4] Liu Q, Huang Z Y, Yin Y, et al. EKT: exercise-aware knowledge tracing for student performance prediction [J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 33(1): 100-115.

[5] Gao W B, Liu Q, Huang Z Y, et al. RCD: relation map driven cognitive diagnosis for intelligent education

systems[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York:ACM,2021:501-510.
[6] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: automated decisions and the GDPR[J]. Harvard Journal of Law & Technology,2018,31(2):841 - 887.
[7] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Demonstrations. San Diego: ACL,2016:97 - 101.
[8] 王静,蒋国平,李汝琦. 反事实推理增强的自适应图对比学习 Transformer 推荐模型[J/OL]. 小型微型计算机系统,2026:1-14[2026-05-09]. https://link.cnki.net/urlid/21.1106.tp.20260129.1058.007.
[9] 谭浩. 结构因果模型及其逻辑表征研究[D]. 湘潭:湘潭大学,2023.
[10] Liu Y J, Zhang T C, Wang X C, et al. New development of cognitive diagnosis models [J]. Frontiers of Computer Science,2023,17(1):171604.
[11] Lundberg S M, Lee S I. A unified approach to interpreting model predictions[C]//Advances in Neural Information Processing Systems. Red Hook: Curran Associates Inc.,2017,30:4768-4777.
[12] He L L, Xiao L, Wang P C, et al. Integrating fine-grained attention into multi-task learning for knowledge tracing[J]. World Wide Web,2023,26(5):3347-3372.
[13] Nakagawa H, Iwasawa Y, Matsuo Y. Graph-based knowledge tracing: modeling student proficiency using graph neural network [C]//Proceedings of the 13st International Conference on Web Intellingence Thessaloniki: IEEE,2019:156-163.
[14] Jain S, Wallace B C. Attention is not explanation[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1. Stroudsburg: ACL,2019:3543-3551.
[15] Yang S S, Duan X W, Yu X S, et al. Debiased cognitive diagnosis: a contrastive counterfactual modeling method via variational autoencoder [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press,2026,40(33):27611-27620.
[16] Luo H S, Zhuang F Z, Xie R, et al. A survey on causal inference for recommendation [J]. The Innovation, 2024,5(2):100590.
[17] Zhang J, Hao B, Chen B, et al. Hierarchical reinforcement learning for course recommendation in MOOCs[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press,2019,33:435-442.
[18] Wu Y K, Lu Z, Wu X T. Counterfactual fairness: unidentification, bound and algorithm[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao, China: ijcai. org, 2019: 1438-1444.
[19] Zhang Y C, Tian J H, Sun J, et al. HKGAT: heterogeneous knowledge graph attention network for explainable recommendation system[J]. Applied Intelligence,2025, 55(6):549.
[20] Feng M, Heffernan N T, Koedinger K R. Addressing the assessment challenge with an online system that tutors as it assesses [J]. User Modeling and User-Adapted Interaction,2009,19(3):243-266.

(上接第 69 页)

参考文献：

[1] 交通运输部公路科学研究院. 隧道沥青路面薄层罩面技术指南[R]. 北京:人民交通出版社,2021.
[2] 姜军. 薄层罩面高弹沥青混合料抗反射裂缝性能试验研究[J]. 公路交通技术,2025,41(4):20-25.
[3] 李运华,李珍,原华. 基于 NovaChip 超薄磨耗层的高速公路沥青路面养护技术应用研究[J]. 公路工程,2019, 44(5):156-161.
[4] 刘好,刘超,刘庚. 超薄磨耗层 NovaChip 层间抗剪强度试验研究[J]. 中外公路,2012,32(5):83-85.
[5] 周勇,邓星鹤,许新权,等. HET-Y(C)型沥青混凝土在提升隧道路面质量中的关键工艺研究[J]. 北方交通, 2024(8):74-77,82.
[6] 朱玺. 木质素纤维对透水沥青混合料路用性能的影响研究[J]. 交通世界,2025(9):45-47.
[7] 吴寅,邱彬,闫朝杰,等. 隧道复杂环境下沥青薄层罩面抗滑耐久性研究[J]. 森林工程,2024,40(2):188-197.
[8] 季节,于淼章,王玉果,等. 沥青混合料工作性能研究综述[J]. 中国公路学报,2024,37(9):170-185.
[9] 崔亚楠,赵琳,孙广宁. 两种沥青胶浆的疲劳及自愈合性能试验[J]. 复合材料学报,2018,35(3):699-705.
[10] Sukhija M, Coleri E. A review on the incorporation of reclaimed asphalt pavement material in asphalt pavements: management practices and strategic techniques[J]. Road Materials and Pavement Design, 2025,26(12):2991-3030.
[11] Luo D, Khater A, Yue Y, et al. The performance of asphalt mixtures modified with lignin fiber and glass fiber: a review [J]. Construction and Building Materials,2019,209:377-387.
[12] 葛冬冬,姜向阳,吕松涛,等. 干法橡胶沥青混合料研究综述:材料、机理、设计与性能[J]. 交通运输工程学报,2025,25(4):1-27.