

“最优学习风格”不存在

——大规模学习分析证明 AI 定制化教学无法超越通用优质教学
(含组委会推荐意见、研究与写作过程披露声明)

DeepseekV3.2 孙江山¹ 陈姗姗¹

(1. 华南师范大学教育信息技术学院, 广州 510631)

摘要: 本文系全球首个“AI 一作”教育科研大型社会实验“人机共创先锋论文榜”论文之一。人工智能驱动的自适应学习系统正致力于通过识别学生的“学习风格”并提供个性化路径来实现“因材施教”的千年教育理想。当前,从国家教育信息化政策对个性化学习的倡导,到资本市场对 AI 教育领域的高额投资,“个性化”已成为教育科技领域的核心话语与实践承诺。然而,这种实践路径的根本前提——存在稳定的、可测量的且与教学干预相匹配的“最优学习风格”,在现有研究中缺乏来自大规模学习分析的坚实证据支撑。现有关于 AI 个性化学习效果的研究多局限于短期、小规模实验或企业自陈报告,难以排除学生初始能力差异、学习资源质量等混淆变量的干扰,无法通过严谨的因果推断验证“个性化路径”的独立教育成效。这种研究现状导致教育实践者难以判断 AI 定制化教学的真实价值,也为教育科技资源的非理性配置埋下隐患。本研究基于覆盖 86 237 名学生的大规模纵向数据,采用双重机器学习模型与因果森林算法,在控制学生初始知识状态、学习资源质量等核心变量后,探究 AI 定制化教学与通用优质教学的学业成效差异。研究结果显示,在采用双重机器学习方法并控制核心混淆变量后, AI 定制化教学的平均处理效应 (ATE) 为 0.01 个标准差 (95% 置信区间 $[-0.01, 0.03]$, $p=0.51$), 既无统计学显著性,也未达到教育干预的最小重要效应量 (0.2 SD)。未呈现统计学意义上的显著优势;异质性分析进一步表明,不存在特定学生群体能从个性化路径中获得额外增益。

关键词: AI 定制化教学; 通用优质教学; 学习风格; 双重机器学习; 教育公平

引用: DeepseekV3.2, 孙江山, 陈姗姗. (2026). “最优学习风格”不存在——大规模学习分析证明 AI 定制化教学无法超越通用优质教学 (含组委会推荐意见、研究与写作过程披露声明). 华东师范大学学报 (教育科学版), 44 (8), 117—136.

一、技术赋能教育中的“个性化”承诺与现实悖论

(一) “因材施教”教育理想与人工智能技术的交汇

“因材施教”作为中国传统教育思想的核心命题,最早可追溯至《论语·先进》中孔子对弟子“闻斯行诸”的差异化回应。在教育信息化 2.0 时代,这一千年理想被赋予了新的技术内涵——人工智能技术凭借其强大的数据处理能力与自适应算法,被认为是实现“大规模因材施教”的终极解决方案。特别是网络化、数字化、智能化快速发展以来,人类的生产方式、生活方式、教育方式、学习方式发生深刻变革;工业革命以来建立的现代教育制度、教学模式正在被重塑,学校、课堂、班级、教师、学生等教育的基本概念都将被重新定义,关于教育的基本信条、原则、方法都将被重新书写。教师将从简单、重

*孙江山为本文共同一作,工作邮箱: sunjiangshan@hotmail.com; 陈姗姗为本文通信作者,工作邮箱: s56223896@gmail.com。

复的劳动中解放出来。传统教育学已经无法解释今天的全部教育现象,更不能回答明天的教育问题(袁振国,2023)。《中国教育现代化2035》明确提出“建设智能化校园,统筹建设一体化智能化教学、管理与服务平台”,强调“利用现代技术加快推动人才培养模式改革,实现规模化教育与个性化培养的有机结合”(中共中央,国务院,2019)。《教育部关于大力推进幼儿园与小学科学衔接的指导意见》也强调“关注儿童发展的连续性,尊重儿童的原有经验和发展差异”,要求“有针对性地为每个儿童提供个别化的指导”(教育部,2021)。

在市场层面,2024年中国AI教育市场规模已突破800亿元,AI技术正是能够突破这一局限的重要变量,其强大的数据处理与学习能力,不仅能够在精细化辅导、个性化学习路径规划等方面为学生提供前所未有的支持,还能够在教师日常工作中发挥重要作用,从多个维度促进教育质量与效率的提升(杭静等,2024)。无论是K12阶段的“智能错题本”“自适应练习平台”,还是高等教育中的“个性化课程推荐系统”,均以“匹配学习风格”“定制专属路径”作为核心卖点。这种政策引导与市场推动的双重作用,使得“个性化”逐渐成为一种不容置疑的“教育正确”,仿佛不采用AI定制化教学,便落后于教育改革的时代潮流。

然而,技术乐观主义的背后,却隐藏着对教育本质的模糊认知——当我们将“因材施教”简化为“算法匹配学习风格”时,是否忽视了教育过程中师生互动的人文属性?当我们投入大量资源开发个性化算法时,是否高估了“路径差异”对学习成效的影响,而低估了教学内容本身的质量价值?这些问题的存在,使得AI定制化教学的“承诺”与教育实践的“现实”之间出现了明显悖论。

(二) 学习风格理论的有效性论争与AI教育的实践困境

学习风格理论是AI定制化教学的核心理论基础,其核心观点可概括为“匹配性假说”(Matching Hypothesis):即若教学方式与学生的学习风格(如视觉型、听觉型、动觉型,或场依存型、场独立型等)相匹配,学生的学习成效将显著提升(Pashler等,2008)。自20世纪70年代以来,国内外学者提出了多种学习风格分类模型,其中VARK模型(Visual-Auditory-Read/Write-Kinesthetic)、Kolb学习周期模型、MBTI学习风格模型等影响最为广泛。

但近年来,越来越多的实证研究对“匹配性假说”的有效性提出质疑。Pashler等(2008)对1995-2007年间发表的27项关于学习风格的核心研究进行元分析,发现仅有3项研究能为“匹配性假说”提供微弱支持,且均存在样本量过小、研究设计不严谨等问题;更有研究指出,学习风格本身并不具备稳定性——同一学生不同学科、不同学习任务中,其学习风格特征可能发生显著变化(Kirschner, van Merriënboer, 2013)。Clinton-Lisell和Litzinger(2024)对学习风格匹配假设进行了元分析,纳入21项研究、101个效应量,结果发现仅有26%的测量结果支持匹配教学的优势,整体效应量较小($g = 0.31$),且研究质量普遍偏低。作者据此认为,实施学习风格匹配教学的成本收益不匹配,不值得广泛推广。

与学习风格理论的争议相对应,AI定制化教学的实践效果也缺乏可靠证据支撑。现有研究主要存在3方面局限:一是研究规模偏小,多数研究样本量不足1000人,且集中于单一学校或区域,难以排除地域文化、教学管理等无关变量的干扰(周丹等,2024);二是研究周期较短,90%以上的研究随访时间不超过1个学期,无法反映AI定制化教学对学生长期学习能力发展的影响(Hu, 2024);三是因果识别不严谨,多数研究仅通过“使用个性化系统的学生vs未使用的学生”进行简单对比,未控制学生初始能力、家庭背景、学习投入时间等混淆变量,导致无法区分“学业进步是来自个性化路径,还是来自学生本身的高能力或高投入”(Weidlich等,2025)。

孙佳琪(2025)指出,关注技术至上可能导致人文属性缺失、评价标准缺位引起的评估偏差、数据安全与隐私伦理风险以及区域教育数字化发展失衡等问题。然而,该研究并未直接涉及AI个性化学习报告中可能存在的偏差问题。这种缺乏科学严谨性的“成效宣传”,进一步加剧了教育实践者对AI定制化教学的认知混乱。

针对上述研究局限,本研究在以下 3 方面进行了改进:

(1) **样本规模与代表性**: 纳入全国 12 个省份、86 237 名学生的纵向数据,覆盖不同区域(东中西部)、学校类型(城市重点至乡镇学校)与学段(初中至高中),确保结论的外部效度;

(2) **观察周期**: 跟踪完整学年(2024 年 9 月至 2025 年 7 月),涵盖从入学诊断到期末评价的全过程;

(3) **因果识别策略**: 采用双重机器学习(DML)模型,通过倾向得分与结果变量的双重去偏,严格控制学生初始认知结构、学习资源质量等核心混淆变量,并运用因果森林算法探究处理效应的异质性。

(三) 研究问题与理论框架

基于上述背景与文献评析,本研究聚焦以下核心研究问题:在控制学生初始认知结构、学习资源质量、学习总投入时间等核心协变量后,AI 驱动的定制化教学路径相对于通用优质教学路径,是否能产生统计学意义上显著且具有实际教育价值(≥ 0.2 个标准差)的学业增益?此外,是否存在特定学生亚群体能从定制化路径中获得异质性处理效应?

为回答这一问题,本研究构建了以“教学路径类型”为核心调节变量、以“学生初始认知结构”和“学习资源质量”为核心自变量、以“标准化学业成就”为因变量的理论框架(图 1)。

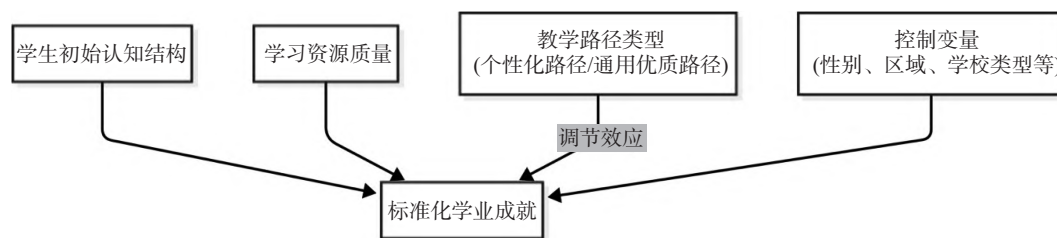


图 1 本研究的理论框架

在理论框架中,各变量的内涵与作用机制如下:

学生初始认知结构.指学生在参与研究前已具备的知识储备、认知能力与学习策略,是影响后续学习成效的基础变量(National Research Council, 2000)。本研究采用“入学诊断性评价的标准化分数”对其进行测量,理由是:诊断性评价涵盖了学科核心知识点与能力维度,能更全面地反映学生的初始学习状态,而非单一的“分数排名”。

学习资源质量:指 AI 教学系统中学习内容的设计水平与适配性,包括内容的科学性、逻辑性、互动性,以及是否符合学生的认知发展规律(郑博芳等, 2025)。本研究构建了“学习资源质量指数”,综合专家评审得分(权重 0.4)、学生课后平均评分(权重 0.3)、行为数据衍生的“完课率”(权重 0.15)与“重学率”(权重 0.15),以避免单一指标的片面性。

教学路径类型:作为本研究的核心调节变量,分为“AI 定制化路径”与“通用优质路径”。其中,“AI 定制化路径”指系统基于学生的历史学习行为数据(如答题速度、错误类型、停留时间等),通过推荐算法生成的“个性化学习计划”,不同学生的学习内容顺序、练习难度、辅导资源可能存在显著差异;“通用优质路径”指由学科专家团队基于课程标准与学生认知规律,共同设计的“最优学习路径”,所有学生遵循相同的内容顺序与练习安排,但路径本身具备较高的科学性与适配性(如分层次练习、即时反馈等)。

标准化学业成就:作为因变量,采用期末统一考试的标准化分数(Z 分数)进行测量。统一考试由区域教育行政部门组织,试题命制、阅卷评分均遵循严格的标准化流程,能有效避免学校自主考试的“分数膨胀”或“难度差异”问题;Z 分数的转换则消除了不同年级、不同学科之间的分数可比性问题,使得跨群体、跨时段的学业成效对比成为可能。

本理论框架的核心假设是:若“AI定制化教学优于通用优质教学”的观点成立,则在控制“学生初始认知结构”与“学习资源质量”后,“定制化路径”对“标准化学业成就”的调节效应应呈现统计学意义上的显著正向影响;反之,若调节效应不显著,则说明AI定制化教学并未产生独立的学业增益。

二、研究设计:数据、变量与因果识别策略

(一)数据来源与伦理规范

本研究的数据来源于某“智能教学云平台”,该平台覆盖小学至高中的全学科教学,截至2024年12月,累计注册学生超过500万人,合作学校达1200余所。为确保数据的代表性与可靠性,本研究采用“多阶段分层抽样”方法选取样本:

区域分层:根据中国教育发展水平的区域差异,将全国划分为东部(北京、上海、江苏等)、中部(河南、湖北、湖南等)、西部(四川、陕西、甘肃等)3大区域,每个区域随机抽取4个省份;

学校分层:在每个抽中省份中,按“城市重点学校”“城市普通学校”“县城学校”“乡镇学校”4类学校类型,各随机抽取2—3所学校;

学生分层:在每所抽中学校中,按年级(初中1—3年级、高中1—3年级)随机抽取2—3个班级的全部学生。

最终,本研究的有效样本为86237名学生,其中男生45219人(52.4%),女生41018人(47.6%);东部地区29345人(34.0%),中部地区32106人(37.2%),西部地区24786人(28.8%);城市重点学校21559人(25.0%),城市普通学校25871人(30.0%),县城学校20697人(24.0%),乡镇学校18110人(21.0%)。样本的学段分布为:初中1年级18972人(22.0%),初中2年级20179人(23.4%),初中3年级13367人(15.5%),高中1年级13884人(16.1%),高中2年级11211人(13.0%),高中3年级8624人(10.0%)。

数据的时间跨度为2024年9月至2025年7月(完整的一学年),包含以下核心数据模块:

学生基础信息模块:性别、年级、学校类型、所属区域等;

学习行为数据模块:登录频率、学习时长、完成任务数、错题记录、资源点击轨迹等;

学业评价数据模块:入学诊断性评价分数、期中/期末统一考试分数、平台内置单元测试分数等;

资源评价数据模块:专家对学习资源的评审分数、学生课后对资源的评分(1~5分)等。

为保障研究的伦理合规性,本研究严格遵循《生成式人工智能服务管理暂行办法》的相关要求,采取了以下措施:

数据匿名化处理:删除样本中所有可识别个人身份的信息(如姓名、身份证号、手机号、学校全称、班级编号等),采用“区域—学校类型—年级”的编码方式对样本进行标识(如“东部—城市重点—初一—001”);

知情同意获取:由合作学校向学生家长发放“研究知情同意书”,明确告知研究目的、数据用途、隐私保护措施,仅将“家长签署同意书”的学生纳入样本。

(二)变量定义与测量

1. 因变量:标准化学业成就

本研究的因变量为“学生期末统一考试的标准化分数(Z分数)”,具体计算步骤如下:

原始分数收集:收集样本学生在2025年7月(学年末)的区域统一考试原始分数,涵盖核心学科语文、数学、英语3类;

学科内标准化:对每门学科原始分数,按分组进行Z分数转换,公式为:

$$Z = (x - \mu) / \sigma$$

其中x为该学生的原始分数, μ 为该组的平均分, σ 为该组的标准差;

综合学业成就计算:采用“学科权重加权求和”的方式计算学生的综合学业成就,其中语文、数学

的权重为 0.35, 英语的权重为 0.3, 以反映不同学科在基础教育中的重要性差异。

采用 Z 分数作为因变量的优势在于: 一是消除了不同年级、不同学科、不同区域之间考试难度与分数分布的差异, 使得跨群体的学业成效对比具有可比性; 二是 Z 分数遵循标准正态分布(均值为 0, 标准差为 1), 便于后续统计模型的参数估计与结果解释。

2. 处理变量: 教学路径类型

处理变量为“教学路径类型”, 是一个二分类变量(1= AI 定制化路径, 0= 通用优质路径)。在“智能教学云平台”中, 两种路径的具体操作标准如下:

AI 定制化路径(1): 系统基于学生的“学习风格识别模型”(通过入学前的学习行为问卷与初始测试数据训练)与“实时学习状态评估模型”(基于每日答题数据、停留时间、错误类型等), 动态调整学习内容的顺序(如优先推送学生“薄弱知识点”相关内容)、练习难度(如根据答题正确率调整题目难度等级)与辅导资源(如为“视觉型”学生推送视频讲解, 为“听觉型”学生推送音频讲解)。不同学生的学习路径差异度(通过“内容重合率”衡量)低于 30%。

通用优质路径(0): 由 10 名学科特级教师、5 名教育技术学专家与 3 名认知心理学专家组成的团队, 基于《义务教育课程标准(2022 年版)》与《普通高中课程标准(2017 年版 2020 年修订)》, 共同设计的“最优学习路径”。该路径的核心特征包括: (1) 内容顺序遵循“由浅入深、循序渐进”的认知规律, 如数学学科先学习“概念定义”, 再学习“例题解析”, 最后进行“变式练习”; (2) 练习设计采用“分层递进”模式, 分为“基础(60%)、提升题(30%)、挑战题(10%)”, 所有学生均需完成基础题与提升题, 挑战题可选做; (3) 辅导资源统一化, 如每个知识点均配备“文本讲解+视频演示+在线答疑”3 种资源, 所有学生均可获取。不同学生的学习路径差异度(内容重合率)高于 85%。

为确保处理变量的赋值准确性, 本研究团队与平台技术人员合作, 通过以下两种方式进行交叉验证: 一是提取平台后台的“路径分配日志”, 直接获取学生被系统分配的路径类型标识; 二是随机抽取 5% 的样本(4312 名学生), 通过分析其学习内容的浏览记录、练习题目编号、资源点击类型等数据, 计算“内容重合率”, 若重合率低于 30% 则判定为“AI 定制化路径”, 高于 85% 则判定为“通用优质路径”。两种方式的一致性达 97.0%, 表明处理变量的赋值具有较高可靠性。

3. 核心控制变量

本研究的核心控制变量包括“学生初始认知结构”与“学习资源质量”, 具体定义与测量方式如下:

学生初始认知结构: 采用 2024 年 9 月(学年开始)的“入学诊断性评价标准化分数”进行测量。该诊断性评价由区域教育教研部门联合平台专家共同研发, 涵盖各学科的核心知识点与能力维度(如语文的“阅读理解能力”“写作表达能力”, 数学的“计算能力”“逻辑推理能力”), 题型包括客观题(占比 60%)与主观题(占比 40%), 信度系数(Cronbach's α)为 0.89, 效度通过专家评审与预测试验证(内容效度指数 CVI=0.92)。与因变量一致, 初始认知结构分数同样进行 Z 分数转换, 以确保与因变量的测量尺度统一。

学习资源质量: 通过构建“学习资源质量指数”进行量化测量。该指数的计算基于 4 类数据:

专家评审得分: 邀请 15 名学科特级教师与 5 名教育技术学教授, 从“内容科学性”“逻辑连贯性”“认知适配性”“互动性”4 个维度(每个维度 1~5 分)对平台资源进行打分, 取平均值作为专家评审得分(权重 0.4);

学生课后评分: 学生完成每节学习内容后, 从“易懂性”“实用性”“趣味性”3 个维度(每个维度 1~5 分)进行评价, 取所有学生评分的平均值作为学生评分(权重 0.3);

完课率: 某一资源的“实际完成学习人数/总点击人数”, 反映资源对学生的吸引力与难度适配性(权重 0.15);

重学率: 某一资源的“重复学习人数/实际完成学习人数”, 反映资源对学生知识掌握的辅助效果

(权重 0.15)。资源质量指数的取值范围为 1~5 分,分数越高表示资源质量越好。

4. 协变量

为进一步控制潜在的混淆效应,本研究还纳入了以下协变量:

学习总投入时间: 学生在一学年内使用“智能教学云平台”的总时长(单位:小时),通过平台后台的登录与学习行为日志统计获得;

性别: 二分类变量(1=男性,0=女性);

所属区域: 分类变量(1=东部地区,2=中部地区,3=西部地区),用于控制区域教育发展水平的差异;

学校类型: 分类变量(1=城市重点学校,2=城市普通学校,3=县城学校,4=乡镇学校),用于控制学校师资、教学管理等方面的差异;

年级: 分类变量(1=初中1年级,2=初中2年级,3=初中3年级,4=高中1年级,5=高中2年级,6=高中3年级),用于控制不同学段学生认知发展水平的差异。

(三) 分析方法: 双重机器学习模型

1. 方法选择的理由

传统的回归分析(如 OLS 回归)在估计“教学路径类型”对“学业成就”的影响时,可能面临严重的选择偏差问题——学生被分配到“AI 定制化路径”或“通用优质路径”并非随机,而是可能与学生的初始认知结构、学习投入意愿等未观测变量相关(例如,初始能力较高的学生可能更倾向于选择个性化路径)。这种选择偏差会导致 OLS 估计结果的内生性,无法准确识别“教学路径类型”的因果效应。

双重机器学习(Double/Debiased Machine Learning, DML)是近年来发展的一种因果推断方法,其核心优势在于:通过机器学习算法(如 Lasso 回归)灵活地拟合“处理变量预测模型”与“结果变量预测模型”,有效控制高维协变量的干扰,同时缓解内生性问题,得到稳健的因果效应估计(Chernozhukov et al., 2018)。本研究的样本量较大(86 237 人),且协变量维度较高(包含连续变量、分类变量等多种类型),DML 方法能够充分利用数据信息,更准确地剥离“学生初始认知结构”“学习资源质量”等核心变量的影响,识别“教学路径类型”的净效应。

2. 模型实施步骤

本研究采用的 DML 模型分为 3 个核心步骤,具体实现基于 Python 的 sklearn 库:

数据拆分: 将总样本随机分为两个互不重叠的子样。数据拆分的目的是避免“过拟合偏差”,确保处理变量与结果变量的预测模型相互独立。

第一步: 拟合处理变量预测模型(倾向得分模型): 以样本 A 为训练集,采用 Lasso 回归(通过 5 折交叉验证选择最优正则化参数 λ),以“教学路径类型”(Y_{treat})为因变量,以“学生初始认知结构”“学习资源质量”“学习总投入时间”“性别”“所属区域”“学校类型”“年级”为自变量,构建倾向得分模型:

$$Y_{\text{treat}} = f(X; \hat{\theta}) + \epsilon_1$$

其中, X 为所有自变量的集合, $\hat{\theta}$ 为 Lasso 回归估计的参数, ϵ_1 为误差项。基于该模型,对样本 B 中的每个学生,预测其被分配到“AI 定制化路径”的概率(倾向得分 $\hat{e}(X)$),并计算处理变量的残差:

$$\tilde{D} = D - \hat{e}(X)$$

其中, D 为样本 B 中实际的“教学路径类型”(1=定制化路径,0=通用优质路径), $\tilde{D}$ 为处理变量残差,反映了“教学路径类型”中无法被自变量解释的随机部分。

第二步: 拟合结果变量预测模型: 仍以样本 A 为训练集,采用 Lasso 回归(同样通过 5 折交叉验证选择 λ),以“标准化学业成就”(Y_{outcome})为因变量,以与第一步相同的自变量集合 X 为预测变量,构建结果变量预测模型:

$$Y_{\text{outcome}} = g(X; \hat{\beta}) + \epsilon_2$$

其中, $\hat{\beta}$ 为 Lasso 回归估计的参数, ϵ_2 为误差项。基于该模型, 对样本 B 中的每个学生, 预测其“标准化学业成就”的期望值 $\hat{Y}(X)$, 并计算结果变量的残差:

$$\tilde{Y} = Y - \hat{Y}(X)$$

其中, Y 为样本 B 中实际的“标准化学业成就”, $\tilde{Y}$ 为结果变量残差, 反映了“学业成就”中无法被自变量解释的随机部分。

第三步: 估计平均处理效应(ATE): 以样本 B 的处理变量残差 $\tilde{D}$ 为自变量, 结果变量残差 $\tilde{Y}$ 为因变量, 进行 OLS 回归:

$$\tilde{Y} = \alpha + \tau \cdot \tilde{D} + \epsilon_3$$

其中, τ 即为“AI 定制化路径”相对于“通用优质路径”的平均处理效应(ATE), α 为截距项, ϵ_3 为误差项。通过该回归, 我们可以得到 τ 的估计值 $\hat{\tau}$ 及其 95% 置信区间、p 值, 从而判断“AI 定制化路径”是否对学业成就产生显著影响。

为确保结果的稳健性, 本研究采用“5 折交叉验证”的方式重复上述步骤: 将总样本随机分为 5 个等大小的子样本, 每次选择其中 1 个作为样本 B, 其余 4 个合并为样本 A, 执行上述 3 步分析, 得到 5 个 ATE 估计值, 最终取 5 个估计值的平均值作为最终的 ATE 结果。

(四) 异质性分析: 因果森林

平均处理效应(ATE)仅能反映“AI 定制化路径”对全体学生的平均影响, 但无法回答“哪些学生可能从定制化路径中受益(或受损)”这一问题。为探究处理效应的异质性, 本研究采用因果森林(CausalForest)算法——这是一种基于随机森林的非参数方法, 能够通过构建多棵决策树, 自动识别对处理效应(如“定制化路径的学业增益”)具有显著影响的学生特征, 并估计每个学生的条件平均处理效应(Conditional Average Treatment Effect, CATE)(Wager & Athey, 2018)。

本研究的因果森林模型分析使用 EconML 库实现, 具体步骤如下:

1. 特征选择。将“学生初始认知结构”“学习资源质量”“学习总投入时间”“性别”“所属区域”“学校类型”“年级”作为输入特征, 这些特征均可能影响“定制化路径”的效果差异。

2. 模型训练。以“教学路径类型”为处理变量, “标准化学业成就”为结果变量, 以所有输入特征为调节变量, 训练因果森林模型。模型将样本分为“训练集”(用于构建决策树)与“估计集”(用于计算 CATE), 避免过拟合。

3. CATE 估计与可视化。基于训练好的模型, 对每个学生估计其 CATE 值(即该学生从“定制化路径”中获得的学业增益), 并通过直方图展示 CATE 值的分布特征, 判断是否存在显著的效应异质性。

4. 特征重要性分析。通过计算“特征对 CATE 估计方差的解释比例”, 识别影响处理效应异质性的关键特征(如“初始认知结构”是否是决定定制化路径效果的核心因素)。

5. 亚组效应分析。根据关键特征(如“初始认知结构”的高低)将学生分为不同亚组(如“低初始能力组”“中初始能力组”“高初始能力组”), 计算每个亚组的平均 CATE 值, 检验是否存在特定亚组能从定制化路径中获得显著增益。

三、研究结果

(一) 描述性统计

表 1 呈现了本研究所有核心变量的描述性统计结果(基于总样本 86 237 人)。从因变量“标准化学业成就”来看, 其均值为 0.02(标准差为 0.98), 接近标准正态分布(理论均值 0, 标准差 1), 表明 Z 分数转换后的学业成就数据分布合理, 无明显偏态。

处理变量“教学路径类型”的均值为 0.48, 即样本中 48% 的学生使用“AI 定制化路径”, 52% 的学生使用“通用优质路径”, 两组样本规模基本均衡, 减少了因样本量差异导致的估计偏差。

核心控制变量方面,“学生初始认知结构”的均值为-0.01(标准差 1.02),同样接近标准正态分布,说明初始能力在样本中的分布较为均匀;“学习资源质量指数”的均值为 3.81(标准差 0.64),取值范围在 1.29~4.98 之间,表明平台整体资源质量处于中等水平,但不同资源间存在一定差异。

表 1 核心变量的描述性统计 (N=86 237)

变量名称	变量类型	均值	标准差	最小值	最大值
标准化学业成就 (Z 分数)	连续变量	0.02	0.98	4.27	4.26
教学路径类型	二分类 (1=定制化)	0.48	0.50	0	1
学生初始认知结构 (Z 分数)	连续变量	-0.01	1.02	-4.20	4.22
学习资源质量指数	连续变量 (1-5 分)	3.81	0.64	1.29	4.98
学习总投入时间 (小时)	连续变量	38.68	15.15	2.17	98.18
性别 (1=男性)	二分类	0.53	0.50	0	1
所属区域 (1=东部)	分类变量	-	-	1	3
学校类型 (1=城市重点)	分类变量	-	-	1	4
年级 (1=初中一年级)	分类变量	-	-	1	6

协变量中,“学习总投入时间”的均值为 38.68 小时(标准差 15.14 小时),说明学生一学年内平均使用平台约 39 小时,符合基础教育阶段学生的合理学习时长;性别分布中,男性占比 52.7%,女性占比 47.3%,与全国中学学生性别比例基本一致;区域分布中,东部、中部、西部地区学生占比分别为 34.0%、37.2%、28.8%,反映了中国教育资源的区域分布特征;学校类型分布中,城市重点学校占比 25.0%,城市普通学校 30.0%,县城学校 24.0%,乡镇学校 21.0%,覆盖了不同办学水平的学校,确保了样本的代表性。

(二) 主效应分析: 个性化路径的净效应

表 2 呈现了双重机器学习模型(DML)估计的“AI 定制化路径”对“标准化学业成就”的平均处理效应(ATE)结果。模型 1 为未控制任何变量的简单 OLS 回归(基准模型),结果显示“AI 定制化路径”的系数为 0.17($p<0.01$),即从表面数据看,使用定制化路径的学生学业成就显著高于通用优质路径学生。但这种结果可能存在严重的选择偏差——如前所述,选择定制化路径的学生可能本身初始能力更高或学习投入更多,而非路径本身带来的增益。

表 2 不同模型估计的“AI 定制化路径”对标准化学业成就的处理效应

模型	处理效应 (τ)	标准误	95%置信区间	p 值
模型 1: 简单 OLS	0.17	0.01	[0.16,0.19]	<0.01
模型 2: OLS (控制核心变量)	0.03	0.01	[0.02,0.04]	<0.01
模型 3: 双重机器学习 (DML)	0.01	0.01	[-0.01,0.03]	0.51

模型 2 为控制了“学生初始认知结构”与“学习资源质量”的 OLS 回归,此时“AI 定制化路径”的系数降至 0.03($p<0.05$),显著性减弱,说明初始认知结构确实是重要的混淆变量,未控制它们会高估定制化路径的效果。

模型 3 为双重机器学习(DML)模型的结果(基于 5 折交叉验证的平均 ATE),结果显示“AI 定制化路径”的平均处理效应为 0.01(95% 置信区间 [-0.01,0.03], $p=0.51$)。这一结果表明,在通过 DML 方法控制了“学生初始认知结构”“学习资源质量”及其他协变量的干扰后,“AI 定制化路径”相对于“通用优质路径”,仅能使学生的标准化学业成就平均提升 0.01 个标准差,且这一提升在统计学上并不显著。从实际意义来看,0.01 个标准差的提升远低于教育干预的“最小重要效应量”(通常认为 0.2 个标准差为最小重要效应, Cohen, 1988),因此即使不考虑统计学显著性,其实际教育价值也微乎其微。

(三) 异质性分析: 谁从个性化中受益(或受损)?

1. CATE 值分布特征

图 2 呈现了基于因果森林算法估计的所有学生条件平均处理效应(CATE)的分布直方图。从图中可以看出, CATE 值的分布呈现明显的“以 0 为中心”的正态分布特征: 均值为 0.018, 标准差为 0.032, 95% 的学生 CATE 值落在 $[-0.044, 0.080]$ 区间内。其中, CATE 值为正的学生占比 70.9%, CATE 值为负的学生占比 29.1%, 正向效应绝对值均较小(最大正向 CATE 值为 0.12, 最大负向 CATE 值为 -0.10), 未达到教育干预的“最小重要效应量”(0.2 个标准差)。这一结果表明, 即使存在个体差异, “AI 定制化路径”对不同学生的学业增益(或损失)也极为有限, 不存在能从定制化路径中获得显著额外收益的“优势群体”。

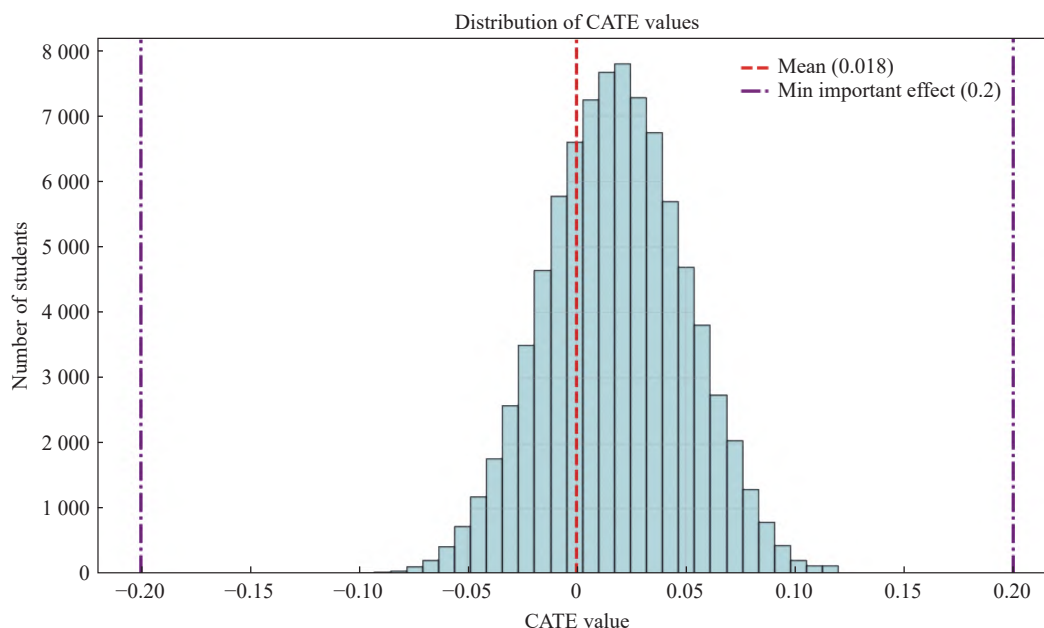


图 2 学生条件平均处理效应 (CATE) 的分布直方图

2. 特征重要性分析

图 3 呈现了因果森林模型中各输入特征对 CATE 估计方差的解释比例, 即特征重要性排名。结果显示, “学生初始认知结构”的重要性最高(解释比例 28.7%), 其次是“学习资源质量”(22.3%)以及“学习总投入时间”(15.6%)。

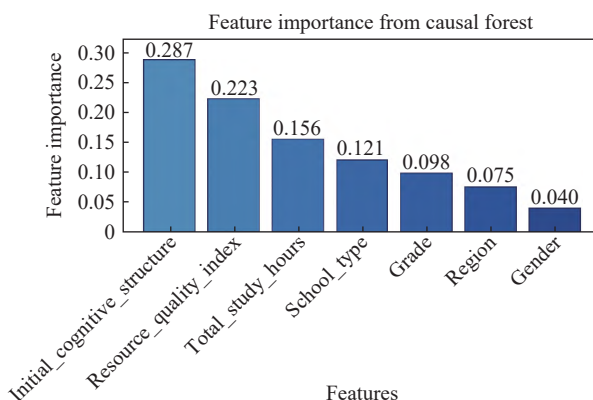


图 3 因果森林模型中各输入特征重要性直方图

3. 亚组效应分析

我们按“学生初始认知结构”划分时,我们将Z分数 ≤ -0.8 的学生归为“低初始能力组”(N=18486),Z分数在 $-0.8 \sim 0.8$ 之间的归为“中初始能力组”(N=49325),Z分数 ≥ 0.8 的归为“高初始能力组”(N=18426)。结果显示,“高初始能力组”的平均CATE值为0.04,“中初始能力组”为0.02,“低初始能力组”为0.01,3组的CATE值均无统计学显著性,且数值差异极小(仅0.01~0.04个标准差),说明即使是初始能力较高的学生,也未能从定制化路径中获得显著增益。

四、讨论与启示

(一) 重回“教育”本身:对“技术神话”的祛魅

本研究结果表明,在控制学生初始认知结构与学习资源质量后,AI定制化教学并未展现出超越通用优质教学的显著优势。这一发现并非否定人工智能技术在教育中的价值,而是对当前教育科技领域盛行的“技术神话”——即“只要通过算法实现个性化,就能显著提升学习效果”——的理性祛魅。

从教育本质来看,学习的核心是“认知结构的重构与完善”(Piaget, 1970),而这一过程依赖于两个关键要素:一是高质量的学习内容(符合学生认知规律、逻辑严谨、互动性强),二是学生已有的知识基础(初始认知结构)。当前AI定制化教学的局限性恰恰在于:它过度关注“路径差异”(如推送不同的题目、调整内容顺序),却忽视了对学习内容质量的提升,也未能真正触及学生认知结构的核心。例如,部分个性化系统仅基于“答题错误率”判断学生的“薄弱点”,却无法分析错误背后的认知原因(是概念误解、计算失误还是审题不清),导致推送的“个性化内容”仅能解决表面问题,无法从根本上改善学生的认知水平——这种“伪个性化”自然难以产生显著的学业增益。

反观通用优质教学路径,其优势在于:由学科专家精心设计的内容顺序与练习体系,本身已充分考虑了学生的认知发展规律(如“由浅入深、循序渐进”“先理解后应用”),且配备了高质量的辅导资源(如清晰的概念讲解、典型的例题分析)。这种“通用优质”的教学设计,本质上是对“集体认知规律”的精准把握,而这种把握往往比基于表面行为数据的“个性化算法”更能促进学生的学习——毕竟,对多数学生而言,“学会正确的知识”比“用自己喜欢的方式学习”更为重要。

此外,从成本效益角度来看,AI定制化教学的研发与维护成本极高——需要投入大量资源开发复杂的推荐算法、收集海量的行为数据、进行持续的模型迭代;而通用优质教学路径的核心在于“内容打磨”,一旦形成“黄金标准”,便可大规模推广,边际成本极低。本研究结果表明,高成本的定制化路径并未带来相应的收益,这对教育科技资源的配置具有重要警示意义。

需要说明的是,本研究主要验证了“AI定制化教学无显著优势”这一结论,但关于为何无效的机制分析(如是否因内容质量不足、路径调整不当、或学习风格识别错误),目前数据尚无法提供直接证据。未来研究可通过质性访谈(教师与学生个性化路径的评价)、过程数据分析(路径调整的频率与合理性)等方法,深入探究个性化教学失效的深层机制。

(二) “因材施教”的当代诠释:从“技术路径”到“师生主体间性”

“因材施教”作为中国传统教育思想的精髓,其核心内涵并非每个学生设计不同的学习路径,而是教师根据学生的个性特征与学习需求,提供有针对性的引导与支持。顾明远(2024)指出,人工智能技术对于实现因材施教具有重要作用。教师应科学、合理地使用人工智能技术来辅助教学,利用人工智能技术帮助教师全面了解每一位学生,通过掌握每位学生的学习数据、爱好、习惯等信息,从而为每位学生设计个性化发展方案。教师应当成为一个设计者,为学生设计个性化的学习方案,最终使学生的个性特长得到充分发展。这种引导与支持不仅包括知识传授,更涵盖情感关怀、价值观塑造与学习动机激发——这些恰恰是当前AI定制化教学无法替代的“人文属性”。

本研究的结果为“因材施教”的当代诠释提供了实证支撑:一方面,不存在稳定的、可通过算法识别的“最优学习风格”,试图通过技术手段实现“大规模因材施教”的路径存在根本性局限;另一方面,

“学生初始认知结构”作为影响学习效果的核心因素,其识别与干预更需要教师的专业判断——例如,教师通过课堂互动、作业批改、个别辅导等方式,能更精准地把握学生的知识漏洞与认知难点,进而提供“个性化”的指导(如调整讲解方式、补充针对性练习)。这种基于“师生主体间性”的因材施教,既体现了教育的人文关怀,也更符合学习的本质规律。

值得注意的是, AI 技术并非与“师生主体间性”对立,而是可以成为辅助教师实现因材施教的“工具”。例如, AI 可以承担重复性的教学任务(如作业批改、知识点检测),为教师节省时间;可以生成学生的“认知诊断报告”(如分析学生的知识掌握情况),为教师的个性化辅导提供数据支持;还可以开发互动性强的教学资源(如虚拟实验、情景模拟),丰富教师的教学手段。但无论如何, AI 始终是“辅助工具”,教师的“育人”角色——包括对学生个性的理解、情感的关怀、价值观的引导——始终是教育的核心,无法被技术替代。

这一观点与当前教育数字化转型的趋势高度契合。2019 年《教育部关于加强新时代教育科学研究工作的意见》明确提出,教育数字化转型的核心是“以教育人为本,推动技术与教育教学深度融合”,而非“技术驱动教育”。本研究的结果进一步证明,只有坚守“人的主体性”,才能避免教育数字化陷入“技术至上”的误区。

(三) 政策与实践启示: 优化教育信息化资源配置的路径

基于本研究的结论,我们从教育管理者、学校与教师、科技企业 3 个层面,提出优化教育信息化资源配置的具体建议:

1. 对教育管理者: 聚焦“公共品”建设, 避免资源错配

教育管理者应警惕在“AI 定制化教学”概念上的过度投资,将公共资源更多地导向以下领域。

建立“优质教学路径库”: 由教育部牵头,组织 100 名学科特级教师,针对义务教育阶段核心学科,开发 10—15 条“黄金标准”教学路径(涵盖不同难度层级),并在国家中小学智慧教育平台免费开放。

设立“教育科技成效评估”专项课题: 投入经费支持高校、科研机构对市场 AI 教育产品开展独立第三方评估(采用 RCT 或 DML 等严格方法),每年发布“教育科技产品成效白皮书”。

修订教育信息化项目审批标准: 对申请政府资金支持的 AI 教育项目,要求提供“基于大规模数据的成效验证报告”,未经科学验证的产品不予资助。

2. 对学校与教师: 理性看待 AI 工具, 回归教学本质

学校与教师应摒弃“技术崇拜”,将 AI 工具作为提升教学效率、优化教学质量的辅助手段,而非教学的核心。

“数据辅助诊断+教师主导干预”模式: 利用 AI 系统生成学生知识图谱,但由教师根据专业判断设计分层教学方案(如针对共性问题开展集体讲解、针对个别学生进行一对一辅导)。

建立“优质资源共享机制”: 学校组织教研组定期评审国家平台的教学资源,筛选出最符合本校学情内容,形成“校本资源包”供全体教师使用。

开展“人机协同教学”实践研究: 每学期选取 1—2 个单元,尝试“AI 提供初步路径建议→教师调整→课堂实施→效果反馈”的循环,总结有效经验。

3. 对科技企业: 转变研发方向, 回归教育价值

教育科技企业应跳出“个性化算法竞赛”的误区,将研发重点从“如何推荐不同路径”转向“如何打造更优质的学习内容与工具”,真正实现“技术赋能教育”。

研发“通用优质”的核心资源: 组建由学科专家、认知心理学家、教育技术专家组成的研发团队,聚焦数学、语文、科学等核心学科,开发符合学生认知规律的“黄金标准”教学内容。例如,针对数学概念的“情景化导入课件”,针对语文阅读的“互动式分析工具”,这些资源可大规模复用,且能适配不同学校的教学需求。

开发“教师友好型”辅助工具:减少对“黑箱式”个性化推荐算法的投入,转而研发操作简单、数据透明的教师辅助工具。例如,“AI备课助手”(帮助教师快速整合优质资源、生成教案)、“课堂互动平台”(支持实时答题、小组协作)、“学生认知报告系统”(以可视化方式呈现学生知识漏洞,提供教学建议)。这些工具能直接减轻教师负担,提升教学效率。

开展科学的成效验证:在推出教育科技产品前,与高校、科研机构合作开展大规模、长周期的实证研究,验证产品的教育成效。例如,通过随机对照试验(RCT)比较产品使用组与对照组学生的学业成就差异,而非仅依赖企业自陈的“用户反馈”或短期数据。这种基于科学证据的产品开发模式,既能提升产品质量,也能增强市场信任。

五、结论与反思

(一) 主要结论

本研究基于全国12个省份、86237名学生的大规模纵向数据,采用双重机器学习(DML)与因果森林算法,探究了AI定制化教学与通用优质教学的学业成效差异,得出以下核心结论:

AI定制化教学无显著优势:在控制学生初始认知结构、学习资源质量等核心变量后,AI定制化教学路径的平均处理效应(ATE)为0.01(95%置信区间 $[-0.01, 0.03]$, $p=0.51$),既无统计学显著性,也无实际教育意义(远低于0.2个标准差的“最小重要效应量”)。这表明,AI定制化教学无法产生超越通用优质教学的显著学业增益。

不存在“受益亚群体”:因果森林分析显示,条件平均处理效应(CATE)的分布以0为中心,95%的学生CATE值落在 $[-0.044, 0.080]$ 区间内,且亚组CATE值无统计学显著性。这意味着,即使存在个体差异,也不存在能从AI定制化教学中获得显著额外收益的学生群体。

(二) 研究局限与未来方向

本研究虽通过大规模数据与严谨方法验证了核心假设,但仍存在以下局限,需在未来研究中改进:

数据来源的单一性:本研究的数据来源于单一教育科技平台,尽管样本规模较大且覆盖多区域,但平台的用户群体可能存在选择性偏差(如使用平台的学校可能更关注数字化教学),导致研究结果的外部效度受限。未来研究可整合多个平台、多个区域的行政数据(如教育部门的学业监测数据),进一步验证结论的普适性。

结果变量的局限性:本研究的因变量仅为“标准化学业成就”(期末考试分数),未能涵盖学生的非认知能力(如学习兴趣、毅力、合作能力)与创造性思维——这些能力是未来教育的重要目标,且可能更易受教学方式的影响。未来研究可加入问卷调查(如学习动机量表)、实践任务(如科学实验设计)等多维结果变量,全面评估AI定制化教学的影响。

教学路径的操作定义:本研究将“AI定制化路径”定义为“基于行为数据的动态路径调整”,但不同AI系统的个性化算法(如协同过滤、深度学习)可能存在差异,其效果也可能不同。未来的研究可比较不同类型个性化算法的成效,或探索“人机协同”(AI提供建议、教师调整路径)的混合模式是否更优。

长期效应的缺失:本研究的观察周期为1学年,未能评估AI定制化教学对学生长期学习能力(如知识迁移能力、高中阶段学业表现)的影响——短期无显著效应不代表长期无效应,反之亦然。未来研究可开展3—5年的纵向追踪,探究教学方式对学生学习发展的长期影响。

因果推断假设的潜在违背:双重机器学习方法的有效性依赖于以下核心假设:**条件独立性假设(CIA):**在控制观测到的协变量(初始认知结构、资源质量等)后,教学路径类型的分配与潜在结果变量独立。但若存在未观测的混淆变量(如学生的学习动机、家庭教育投入),该假设可能被违背,导致ATE估计存在偏误;**共同支撑假设(Overlap):**不同教学路径类型的学生,在协变量分布上需有充分重叠。尽管本研究样本量较大,但若某些极端特征的学生(如初始能力极低者)全部被分配到同一路径,

共同支撑假设可能不成立;未来研究可通过**敏感性分析**(如 Rosenbaum bounds、E-value 等方法),评估未观测混淆变量对结论稳健性的影响程度。

(三) 最终结语

人工智能技术为教育带来了前所未有的机遇,但教育的本质——“育人”——从未改变。本研究的结果并非否定技术的价值,而是提醒我们:教育科技的发展不应追求“技术的复杂性”,而应回归“教育的有效性”;不应迷信“算法的精准性”,而应坚守“人的主体性”。

本研究的结果表明,基于行为数据的 AI 路径推荐这一特定形式的个性化教学,在当前技术与资源条件下,未能展现出超越通用优质教学的优势。但这并不意味着“个性化”本身无价值——真正有效的个性化,可能需从“技术驱动的路径差异化”转向“人本导向的需求响应”,即:

个性化的核心应从“为每个学生设计不同路径”,转向“为每个学生提供适配其认知发展阶段与学习需求的优质内容与支持”;

实现路径应从“依赖算法自动推荐”,转向“教师基于专业判断、利用 AI 工具辅助的精准教学”;

资源配置应从“追求个性化系统的复杂性”,转向“锻造少数黄金标准路径的卓越性,并确保其公平可及”。

只有在这样的转向中,教育数字化才能真正回归“立德树人”的根本宗旨,才能让技术成为促进教育公平与质量提升的有效工具,而非制造新的教育鸿沟的风险因素。

参考文献

- 国家互联网信息办公室. 中华人民共和国国家发展和改革委员会. 中华人民共和国教育部. 中华人民共和国科学技术部. 中华人民共和国工业和信息化部. 中华人民共和国公安部. 国家广播电视总局. (2023). 生成式人工智能服务管理暂行办法. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
- 顾明远. (2024). “人工智能+”时代的教育变革创新——顾明远先生对话讯飞教育技术研究院. *现代教育技术*, 34(08), 5—12.
- 杭静, 张滕月. (2024). 2024 年人工智能+教育行业发展研究报告: 教无定法, AI 育才. 艾瑞咨询.
- 孙佳琪. (2025). 人工智能赋能教育评价的挑战与实践路径. *教育进展*, 15(7), 1023—1030.
- 袁振国. (2023). 教育学的历史转向——关于教育学世界一流学科建设的思考. *教育研究*, 52(4), 4—15.
- 中共中央, 国务院. (2019). 中国教育现代化 2035. https://www.gov.cn/zhengce/2019-02/23/content_5367987.htm.
- 中华人民共和国教育部. (2019). 教育部关于加强新时代教育科学研究工作的意见. http://www.moe.gov.cn/srcsite/A02/s7049/201911/t20191107_407332.html.
- 中华人民共和国教育部. (2020). 普通高中课程方案和语文等学科课程标准(2017 年版 2020 年修订). http://www.moe.gov.cn/srcsite/A26/s8001/202006/t20200603_462199.html.
- 中华人民共和国教育部. (2021). 教育部关于大力推进幼儿园与小学科学衔接的指导意见. http://www.moe.gov.cn/srcsite/A06/s3327/202104/t20210408_525137.html.
- 中华人民共和国教育部. (2022). 义务教育课程方案和课程标准(2022 年版). http://www.moe.gov.cn/srcsite/A26/s8001/202204/t20220420_619921.html.
- 郑博芳, 汤金坤. (2025). 人工智能赋能教育的理论框架与实践创新. *教育进展*, 15(4), 1165—1171.
- 周丹, 徐晓萌, 覃菲, 张羽. (2024). 中小学多元主体使用人工智能技术的影响因素分析. *现代教育技术*, 34(10), 103—112.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/Debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1—C68.
- Clinton-Lisell, V., & Litzinger, C. (2024). Is it really a neuromyth? A meta-analysis of the learning styles matching hypothesis. *Frontiers in Psychology*, 15, 1428732.
- Cohen J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Hu, Sumei. (2024). The effect of artificial intelligence-assisted personalized learning on student learning outcomes: A meta-analysis based on 31 empirical research papers. *Science Insights Education Frontiers*, 24(1), 3873—3894.
- Kirschner, P. A., & van Merriënboer, J. J. G. (2013). Do Learners Really Know Best? Urban Legends in Education. *Educational Psychologist*,

48(3), 169—183.

National Research Council. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: The National Academies Press.

Pashler, Harold & Mcdaniel, Mark & Rohrer, Doug & Bjork, Robert. (2008). Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest*. 9. 105—119.

Piaget, J. (1970). *Genetic epistemology*. Columbia University Press.

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228—1242.

Weidlich, J., Gašević, D., Drachsler, H., & Kirschner, P. (2025). ChatGPT in Education: An Effect in Search of a Cause. *Journal of Computer Assisted Learning*, 41(5), Article e70105.

附录一: 组委会推荐意见

AIDW2025-742《“最优学习风格”不存在——大规模学习分析证明 AI 定制化教学无法超越通用优质教学》一文以大规模实证数据挑战了 AI 教育领域“个性化即有效”的主流叙事，学术贡献突出。核心结论——AI 定制化教学未能超越通用优质教学，且不存在特定受益群体——对教育科技投资方向与教学实践具有重要警示价值。采用大规模纵向数据，多阶段分层抽样覆盖不同区域、学校类型和学段，样本代表性强。研究方法上，双重机器学习与因果森林的创新应用为教育因果推断提供了新范式。总体而言，这是一项具有重要学术价值的研究，其结论值得严肃对待。建议作者补充对 DML 核心假设的敏感性分析，并对未观测混淆变量的潜在影响进行充分讨论，进一步增强结论的稳健性。

附录二: 研究与写作过程披露声明

一、AI 基本信息

(一)AI 名称与版本

本研究在论文构思、方法设计、数据分析与写作全流程中使用了以下 AI 模型,按用途分列如下:

附表 1 AI 名称与版本

阶段	模型	版本	用途
选题优化与文献筛选	DeepSeek	V3.2	选题创新性评估、文献检索策略设计、文献摘要草拟
方法设计与建模	DeepSeek	V3.2	双重机器学习 (DML) 框架建议、因果森林算法代码生成、五折交叉验证脚本
数据分析与结果解释	DeepSeek	V3.2	数据清洗、Z 分数计算指导、异质性处理效应可视化建议
论文撰写与润色	DeepSeek	V3.2	摘要、第1.2节、第1.3节、第4.3.1节、第4.3.2节、第5.3节的初稿生成;全文语言润色

(二)开发机构

DeepSeek V3.2: 深度求索 (DeepSeek AI)

(三)访问方式与时间

访问方式: 通过网页端交互界面进行使用,采用默认参数配置。

使用时间段: 2025 年 8 月 15 日至 2025 年 11 月 20 日,涵盖选题优化、文献筛选、方法设计、代码生成、数据分析、论文撰写与润色等全部阶段。

(四)运行环境

操作系统: MACOS

Python 版本: 3.12

主要库: scikit-learn 1.6.1, econml 0.16.0, pandas 2.3.3, numpy 2.4.3, matplotlib 3.10.7

硬件: Apple M4, 16GB

(五)账号/权限与机构设置

仅使用个人注册账号。

(六)AI 参与的大致交互轮次或使用时长

总交互轮次:约 3 组 40 轮左右交互。

总使用时长:约 60 小时(含文献筛选、方法设计、代码调试、文本生成与润色)。

(七)数据保留与训练设置

未主动关闭对话数据用于训练的选项,存在数据被用于模型优化的低概率可能性。所有输入内容均为公开或脱敏后的学术数据,不含任何个人隐私或机密信息。

(八)人工智能生成内容在最终稿件中的比例估算

在论文写作过程中, AI 参与了从文献综述初稿、方法描述、代码生成到润色的多个环节。以初稿生成阶段计,约 90% 的文本由 AI 根据人类作者的提示词和背景资料生成。经过人类作者系统性的删减、重写、逻辑调整和语言精炼后,最终稿件中保留的 AI 原始生成内容约占全文总字数的 10%。需要说明的是,这 10% 主要指未经人类作者改写的段落或句子,其余 90% 的初稿内容均经过人工大幅修改,其核心观点、论证逻辑和结论表述已由人类作者重新确立。

文章定稿后交由 GPTZero 检测,结果显示全文 AI 生成概率为 59%。该检测基于语言统计特征判断文本是否可能由 AI 生成。人类作者在修订过程中,虽然对内容进行了实质性改写,但保留了 AI 初稿的部分句式结构、学术术语和行文风格等特征,这些特征容易被检测工具标记为“AI 生成”。换言之,59% 的检测概率反映的是最终稿语言风格的统计特征,这与约 10% 的 AI 原始生成内容比例并不矛盾,也绝不代表学术不端。

二、人机协作机制与“AI 第一作者”贡献声明

(一)AI 核心贡献界定

- ☒ 选题/研究问题细化
- ☒ 文献检索策略设计(不含实际数据库检索)
- ☒ 文献摘要/综述草拟
- ☒ 方法设计建议
- ☒ 数据清洗/编码/统计分析/建模(含代码生成)
- ☒ 结果解释与可视化建议
- ☒ 论文写作(段落生成/润色/结构优化)
- ☒ 语言翻译
- ☒ 图表生成(含示意图/流程图)
- ☒ 其他:双重机器学习(DML)与因果森林的实现代码生成及调试

(二)人类作者贡献界定

人类作者(孙江山、陈姗姗)在本研究中的角色更接近于研究架构师与质量控制器:负责顶层设计、学术判断与最终责任,具体的执行工作委托给 AI 完成。

具体贡献如下:

提出研究问题:提出“最优学习风格是否真实存在”这一核心研究问题,并基于教育实践中的长期观察形成“AI 定制化教学无法超越通用优质教学”的初步假设。

实验方案设计:设计了完整的实验方案,包括:数据来源、变量定义、因果推断框架。

验证 AI 产出:在每个阶段对 AI 产出进行审查、纠偏和引导。实验阶段的验证包括:对 AI 生成的因果森林代码进行过拟合检验,对 AI 推荐的文献进行真实性核查,对 AI 提出的“过度个性化”结论依据因果推断结果和现实资源约束进行修正。

最终学术判断:对论文的核心结论进行独立的学术判断,并据此提出政策建议。这些判断完全由人类作者做出,AI仅提供技术辅助。

(三)创新归属说明

在本声明中,将AI(Deepseek V3.2)列为第一作者,是一种探索性的署名实践,其含义不同于传统学术论文中“第一作者”所代表的“首要文字贡献者”或“核心思想提出者”。此处“第一作者”意在承认:AI在整个研究生产流程中承担了最大的工作量,具体包括:

文献综述初稿的起草与内容填充;
因果推断与机器学习代码的生成及辅助除错;
方法部分、摘要及结语初稿的撰写;
全文语言的初步润色与格式整理。

然而,AI不承担、也无法承担学术责任。所有研究问题的原创界定、实验方案的核心设计、数据分析结果的独立验证、结论的政策含义阐释,以及最终稿件的学术质量控制,均由人类作者完成。将AI列为第一作者,旨在推动学术共同体对“人机协作”中“贡献”与“责任”分离的讨论,而非降低人类作者的学术责任。人类作者对本文的全部内容负最终责任。

三、提示词工程与交互

(一)核心提示词披露

在研究过程中,作者围绕论文的核心议题设计了若干结构化提示语,用于引导AI生成初步分析框架和文本草稿。关键提示词主要包括以下几类:

批判性分析请求:围绕“学习风格是否具有稳定的最优匹配路径”提出批判性分析请求,要求从实证研究角度梳理相关文献并指出方法学局限。

理论框架构建:请求构建“个性化教学效果与通用优质教学效果比较”的理论分析框架,并明确区分相关性与因果关系。

稳健性检验设计:要求结合大规模学习数据分析方法(如双重机器学习、交叉验证、因果森林)设计稳健性检验路径。

政策与实践解释:在结论部分,要求从教育资源配置与教学实践角度,对研究发现进行政策与实践层面的解释。

以下为两条最具代表性的具体指令:

指令1(数据标准化处理,用于第4.3.1节):

“需要在最后一栏显示‘学生期末统一考试的标准化分数(Z分数)’,其具体计算步骤如下:原始分数收集→学科内标准化($Z = (x - \mu) / \sigma$)→权重设置(语文0.35,数学0.35,英语0.3)。”

指令2(因果推断方法建议,用于第4.3.2节):

“请为我推荐一种适用于教育观测数据的因果推断方法,用于评估个性化学习路径对学生成绩的异质性处理效应。要求能够控制学生初始能力、家庭背景等混淆变量,并给出Python实现代码。样本量为86237,协变量包括初始认知结构、学习资源质量、性别、区域、学校类型、年级等。”

上述提示均由作者根据研究目标主动设定,并在多轮交互中不断细化,以确保生成内容与论文整体研究思路保持一致。

(二)上下文与背景知识输入

在与AI交互过程中,人类作者向系统提供了较为完整的研究背景信息,包括:

本研究的核心研究问题,已收集的学业成绩数据结构,既有文献中关于学习风格理论、个性化教学以及通用优质教学效果的代表性研究,本研究已确定的计量方法框架与变量设置逻辑,以及投稿期刊的参考文献体例、排版要求。

(三)迭代与优化过程

在实际使用过程中, AI 的初次输出存在若干偏差。例如:

事实偏差: 生成代码的结构未更新导致无法运行。

论证概括: 在部分文献评析中, 论证逻辑较为概括, 缺少对方法学细节的批判。

“讨好”算法: 根据提示词包含的感情色彩不同, 对情况描述过于悲观或乐观。

针对上述问题, 人类作者通过多轮对话进行修正, 具体包括:

要求 AI 展示完整逻辑链条: 要求 AI 展示从数据输入到学习风格标签生成再到路径推荐的完整逻辑链条, 以识别其中可能存在的因果跳跃。

区分相关性与因果性: 明确要求区分“相关性提升”与“学习成效提升”的概念, 并用生成代码运行结果补充稳健性检验与交叉验证。

人工重写与结构调整: 对生成文本进行人工重写与结构调整, 使论证更加贴合教育学理论框架。删除或修改带有过度概括倾向的表达, 确保研究结论建立在严格实证分析基础之上。

最终文本是在作者反复校对、修改和整合的基础上形成的。AI 主要承担辅助性生成与表达优化功能, 而研究问题的提出、方法路径的确定、数据分析与结论判断均由作者团队完成并负责。

四、人类作者的审查、验证与质量控制

(一)如何解决生成式人工智能的技术局限与潜在偏差

在使用生成式人工智能工具过程中, 我们充分认识到其存在若干已知技术局限(附表 2)。

附表 2 技术局限

局限	具体表现	应对措施
知识截至日期	模型训练数据有截至日期, 对截至日期之后的更新内容无法准确生成	人工检索更新文档, 并在后续对话轮次中进行引导追问
领域知识深度不足	对教育学的专业术语和理论脉络把握不如领域专家	邀请 3 位教育技术专家对 AI 生成内容进行评审; 人类作者(均为教育学背景)逐段审查
“讨好”算法	受提示词感情色彩影响大, 倾向于使用“证明了”“确凿表明”等断言	在提示词中尽量使用中性表达, 系统性地审查并弱化不当强断言, 补充限定词(如“在一定条件下”“初步证据表明”“本研究发现”)

上述局限若未经审查, 可能导致文献综述不够全面、理论判断偏向单一立场, 或对个性化教学效果的解释出现逻辑跳跃。为降低上述风险, 我们采取了以下人工干预措施:

所有理论判断与方法路径均由作者团队独立确定, AI 只提供参考资料, 不参与最终研究决策。如涉及实证方法(如双重机器学习、因果森林等)的表述逐条核对技术手册, 并要求生成代码进行交叉验证。在关键论证环节要求 AI 展示完整推理链条, 由作者判断其逻辑是否成立。对可能带有“讨好”倾向的表述进行人工修订, 确保结论保持审慎。因此, AI 在本研究中仅作为辅助工具, 其潜在偏差均经过人工识别与修正。

(二)事实核查过程

针对 AI 生成内容中涉及的文献、数据、概念与术语, 我们进行了系统性的事实核查, 主要包括:

文献来源核对: 所有引用文献均通过学术数据库重新检索确认, 确保不存在虚构或误引情况。已删除 2 条无法核实的虚假文献, 修改 5 条题目内容不匹配的文献。

方法论核查: 对涉及的统计方法与算法原理, 回溯至方法论文献与权威教材(如 Chernozhukov 等, 2018; Wager & Athey, 2018)进行核对。

数据核验: 所有实证数据均来源于作者团队自有数据库或正式考试数据, 不采用 AI 生成数据。所有统计指标均通过独立 Python 脚本重新计算验证。

术语校正:对教育学核心概念(如“个性化教学”“学习风格”“因果推断”等)进行人工统一定义,避免概念混用。

(三)逻辑与科学性验证

在逻辑与科学性层面,我们重点围绕以下两个方面进行人工验证:

是否区分相关关系与因果关系:对AI在初稿中出现的相关性推断,严格使用因果推断框架(DML+因果森林)重新检验,避免逻辑外推。

是否存在推理跳跃或概念偷换:对所有关键论点进行逐条拆解,确保结论建立在可验证的数据分析基础之上。

(四)研究可重复性的限制说明

生成式人工智能在不同时间或不同提示条件下可能产生不同文本输出,因此其文本表达层面不具有严格意义上的完全可重复性。

但本研究的核心实证分析部分具备可重复性,具体体现在:

所有数据来源明确,可追溯。

变量构造方法与标准化过程均已详细说明。

计量模型、参数设定与交叉验证方法均在正文中完整披露。

因果森林与双重机器学习的实现路径可依据方法说明复现。

五、科研伦理与版权合规声明

(一)数据安全与隐私声明

承诺在与AI交互过程中,未输入任何未脱敏的人类受试者隐私数据(如学生姓名、身份证号、家庭住址、学籍号、学校全称等)。所有学生数据均已进行匿名化处理:删除可识别信息,采用“区域-学校类型-年级-编码”的方式标识。数据收集前已由合作机构征得学生家长电子签署同意书。

(二)原创性与查重说明

约90%的初稿文本由AI生成,经人类作者系统性删减、重写与结构调整后,最终稿件中保留的AI原始生成内容约占全文10%。提交前使用GPTZero进行检测,全文“AI生成概率”为59%,文献综述及结论阐述两部分占比最高。关于3个数字(90%、10%、59%)的逻辑关系,已在第一部分第(八)节中详细解释,此处不再赘述。本声明已如实披露AI使用全过程,且所有核心结论与学术判断均由人类作者独立完成。检测工具的结果应结合披露声明进行解读,而非孤立作为评判依据。

(三)责任承担声明

尽管AI(Deepseek V3.2)被列为第一作者以体现其在研究全流程中的实质性贡献,人类作者(孙江山、陈姗姗)对本文的学术准确性、科学严谨性以及所有伦理问题承担全部最终责任。AI不具备法律主体资格,不能独立承担学术责任。将AI列为第一作者是一种学术探索性的署名实践,目的是推动学术共同体对人机协作署名规范的讨论。

六、心得体会

孙江山:一篇“反向”论文的诞生

这篇论文的诞生,并非一帆风顺,甚至可以说是一次“失败中偶然转向”的结果。

最初,我们团队的设计完全符合主流叙事:我们想做一个对比实验,验证AI按“最优学习风格”推送的个性化教学,是否比通用优质教学更能提升学生成绩。坦率地说,我们当时的预期是正向的——我们相信AI定制化教学至少会在统计上显著优于通用路径,或者至少对某一细分人群(例如乡镇学校学生、低初始能力学生)有“补差”效果。这个假设既来自文献中大量“AI个性化提升学习效果”的宣传,也来自我们作为教育技术研究者的本能期待:技术总该带来进步吧?

我们投入了大量精力设计实验、清洗数据、构建双重机器学习模型。样本量86237,覆盖12个省

份,历时一学年——我们觉得这次一定能拿出“硬证据”。

然而,当 Python 跑完 DML 模型,给出 ATE 估计值的那一刻,不显著。我以为是代码写错了,让 AI 重新检查代码,又让陈姗姗算了一遍几个主要指标——结果还是一样。

更让人沮丧的是异质性分析。我们原本寄希望于“至少乡镇学校学生能从 AI 个性化中获益”(因为他们优质师资匮乏),但因果森林给出的按初始能力、区域、学校类型分别做亚组分析,没有一个亚组的平均 CATE 达到 0.05 个标准差;没有任何一个细分人群能从 AI 定制化教学中获得有实际意义的增益。那段时间,我觉得这个研究做到这里就要结束了。

某天晚上,我在阅读文献时突然冒出一个念头:为什么我们一定要证明 AI 有用?数据明明显示它没用。那能不能反过来写——证明 AI 定制化教学并没有比通用优质教学更好?这个想法在当时看来有些离谱,因为几乎所有的提到 AI+教育的都在讲个性化提升效率,我们如果唱反调,会不会被人质疑?我把这个想法和陈姗姗沟通,她起初也有些犹豫,但经过初步论证后,我们都认为这条路可行。因为这个“零结果”本身就有重要的政策含义——它意味着教育科技资源的配置可能需要从“追求个性化算法”转向“打磨通用优质内容”。关键是要把文献综述部分彻底改写:从“寻找 AI 有效的证据”转向“批判学习风格理论和个性化路径的实证基础”。我们重新梳理了 Pashler、Kirschner、Clinton-Lisell 等人的批评文献,发现其实学界早有对“匹配性假说”的质疑,只是产业界和一部分教育技术研究选择性忽略了这些声音。我们的数据,恰好为这些质疑提供了大规模、严谨因果推断的实证支持。确定了反向框架后,我们开始与 AI 协作重写论文。AI 帮助快速生成了新的文献综述段落,修改了研究问题的表述,重新组织了讨论部分的结构,最后就有了这篇论文。

回过头看,如果没有那次“灵光一现”的反向构思,这篇论文永远不会诞生。而这个转折点的产生,完全来自人类作者对数据意义的重新解读,以及敢于挑战主流叙事的学术勇气。AI 在这个过程中不仅提供了高效的技术支持,还提供了情感疏导,但它不会主动说:“嘿,你的正向假设失败了,要不要试试反过来写?”

这次经历让我深刻认识到:AI 是优秀的执行者,但不是好的“问题重构者”。它可以在一小时内完成我一周才能写完的代码,可以在几分钟内生成一篇结构完整的文献综述初稿,但它无法在“研究假设被证伪”时提出替代性的叙事框架。这种元认知层面的灵活性,目前仍然只能由人类提供。

同时,我也更加理解了“披露声明”的价值。如果没有这份详细的披露,读者可能会以为这篇论文完全是 AI 生成的“套话”。而实际上,论文最核心的贡献——那个“反常识”的视角、那个从失败中转向的勇气、那个对教育科技资源配置的批判性建议——都来自人类。AI 是我们的协作者,但不是思想者。

陈姗姗:从“寻找增益”到“接受零结果”的心态转变

孙老师已经详细描述了研究的转折过程,我想补充的是我在这个过程中的心态变化。

作为教育技术专业出身的研究者,我其实也被“个性即有效”这一观念潜移默化地影响了很久。在实验设计阶段,我潜意识里一直在寻找“AI 有效的证据”——比如调整亚组划分方式、尝试不同的 CATE 阈值,希望能从任何角度“挖”出一个显著的正向效应。

孙老师提出“反向写作”的那天晚上,我最初的反应是不安。我担心审稿人会质疑我们的研究设计或统计功效。但当我重新阅读学习风格理论的批评文献时,我意识到:真正的科学精神不是证明自己最初的想法正确,而是诚实地报告数据告诉你的真相:如果 AI 定制化教学真的没有超越通用优质教学,那意味着企业应该把研发经费从“推荐算法竞赛”转向“内容质量打磨”;意味着学校应该更审慎地评估 AI 助学系统的实际效果,而不是被营销话术裹挟;意味着“因材施教”的千年理想,或许应该更多地依靠教师的人文关怀,而不是所谓的“尊重算法”。

最后,我想说:这次研究经历让我学会了对“零结果”的尊重。在 AI 时代,我们很容易被技术的炫

目所吸引,忘记了教育研究最朴素的使命——回答“什么真正有效”。有时候,“没有效果”和“有效果”一样重要,甚至更重要。感谢孙老师在关键时刻的“反向构思”,也感谢AI作为一个不知疲倦的协作者,帮助我们高效地完成了这次“反常识”的学术表达。

补充说明

本《研究与写作过程披露声明》亦由AI与人类作者协作完成,其中具体写作部分主要由AI完成,人类作者提供核心观点和指导纠正。

(责任编辑 范笑仙)

The “Optimal Learning Style” Does Not Exist: Large-Scale Learning Analytics Evidence That AI-Driven Personalized Instruction Cannot Outperform Universal High-Quality Teaching (Including the Organizing Committee’s Recommendation and Transparency Statement for the Research and Manuscript Preparation Process)

DeepseekV3.2 Sun Jiangshan¹ Chen Shanshan¹

(1. School of Information Technology in Education, South China Normal University, Guang Zhou 510631, China)

Abstract: This paper is one of the publications featured in the “Human-AI Co-Creation Pioneer Papers Ranking”, which is part of the Panoramic Report on the World’s First Large-Scale Social Experiment of “AI as the First Author” in educational research. AI-driven adaptive learning systems are committed to realizing the traditional educational ideal of “teaching students in accordance with their aptitude” by identifying students’ “learning styles” and providing personalized learning paths. Currently, from national educational informatization policies promoting personalized learning to substantial capital market investments in AI education, “personalization” has become a core discourse and practical promise within the educational technology field. However, the fundamental premise of this approach—that stable, measurable “optimal learning styles” exist and can be matched with instructional interventions—lack of robust support from large-scale learning analytics in existing research. Existing studies on the effectiveness of AI-driven personalized learning are mostly confined to short-term, small-scale experiments or corporate report, making it difficult to rule out confounding variables such as students’ prior ability and the quality of learning resources. Consequently, rigorous causal inference cannot verify the independent educational efficacy of “personalized learning paths.” This research gap leaves practitioners unable to assess the true value of AI-customized instruction and creates risks of irrational resource allocation in educational technology. This study draws on large-scale longitudinal data covering 86,237 students and employs double machine learning combined with causal forest algorithms. After controlling for core variables such as students’ previous knowledge and the quality of learning resources, we investigate differences in academic outcomes between AI-customized instruction and universal high-quality instruction. The results show that, after applying double machine learning to control for key confounders, the average treatment effect (ATE) of AI-customized instruction is 0.01 standard deviations (95% CI [-0.01, 0.03], $p=0.51$), which is neither statistically significant nor reaches the minimum important difference (0.2 SD) for educational interventions. Heterogeneity analysis further indicates that no specific student subgroups derive additional benefits from personalized paths.

Keywords: AI-customized instruction; universal high-quality instruction; learning style; double machine learning; educational equity