

生成式AI辅助 高中数学试题命制的实证研究

——基于人机协同的案例分析报告

◎ 王中苏，张红（重庆市第一中学校，重庆 400030）

• **摘要：**生成式 AI 可提升高中数学试题初稿生成效率，但在边界条件、逻辑严密性与情境真实性方面仍有局限。本研究结合问卷调查与多案例分析，基于 192 份有效教师问卷及四类典型试题的迭代记录，检验“知识—能力—素养”三维评价框架在人机协同命题中的适用性。结果显示，教师总体满意度为 3.41 分，70.98% 的受访者认为提示词工程存在门槛，且多数教师指出 AI 生成内容在知识覆盖与数理严谨性上仍需人工修订。案例分析表明，AI 的偏差主要集中在尺度失真、对象指代含混、离散与连续混用以及样本空间设定不严等方面。据此，本研究提炼出高中数学人机协同命题的流程与教师介入要点，以供相关教育工作者参考。

• **关键词：**生成式 AI；高中数学命题；人机协同；三维评价框架

Empirical Research on Generative AI Assisted High School Mathematics Test Item Construction:

A Case Analysis Based on Human-AI Collaboration

◎ Wang Zhongsu, Zhang Hong (Chongqing No.1 Secondary School, Chongqing 400030)

• **Abstract:** Generative AI enhances efficiency in drafting high school math test items but has limitations in logic, context and boundary conditions. Based on 192 teacher surveys and iterative records of four typical items, the research validates a three-dimensional “knowledge-competency-literacy” framework. Teacher satisfaction scores 3.41, with 70.98% noting prompt engineering barriers and most requiring manual revision. AI deviations include scale distortion, ambiguous references, discretecontinuous confusion, and sample space issues. Therefore the research outlines a collaborative item development process and key teacher intervention points to serve as a reference for educators.

• **Keywords:** Generative AI; High School Mathematics Item Development; Human-AI Collaboration; Three-Dimensional Evaluation Framework

■ 基金项目：重庆市教育科学“十四五”规划 2025 年度规划一般课题“基于生成式人工智能的高中数学试题命制研究”（课题批准号：K25YG1070333）。

一、引言

（一）研究背景与问题

生成式人工智能进入基础教育后，试题命制效率明显提升，但数学命题并不只是语言生成问题，更关乎逻辑闭合、符号规范与条件完整。AI 在生成复杂数学题时，常因情境失真、条件遗漏或推理断裂而偏离命题要求^[1]。人机协同命题的关键，已不在于“能否生成”，而在于“是否可用、是否规范、是否成立”。

（二）研究现状

现有研究大多讨论大语言模型的一般生成能力及教育应用前景^[2]，对高中数学命题这类高度依赖数理逻辑的任务关注仍显不足。尤其是将教师问卷、真实修订过程和学科判断结合起来考察的研究较少。基于此，本研究以问卷数据与案例修订记录为基础，分析 AI 生成试题如何在教师介入下回归数学规范。

（三）研究方法与样本设计

本研究采用问卷调查与案例分析相结合的方法。问卷面向市、区两级重点高中发放 210 份，回收 195 份，剔除无效问卷后得到 192 份有效样本；量表 Cronbach's α 为 0.87，KMO 值为 0.82，信效度较好。案例部分选取解析几何、函数与代数、立体几何、概率统计四类题目，记录人机协同中的关键修订环节。

二、理论框架与技术路径

（一）三维评价框架的学理构建

本研究提出“知识—能力—素养”三维评价框架。就命题实践而言，AI 较擅长将现实材料转写为数学语言，即完成表层的“水平数学化”；而真正困难的是让条件、结论、推理与限制在数学系统内部形成闭合，这对应更深层的“垂直数学化”。该框架的作用，不在于增加概念包装，而在于帮助教师迅速识别命题的问题究竟出在考点选择、推理结构还是情境设计^[3]。

表 1 三维评价框架的操作化指标

维度	核心要求	AI 常见缺陷与表现
知识 维度	考点准确、表述清楚、无超纲内容	知识覆盖有偏差， 隐藏条件容易漏掉
能力 维度	推理链条完整，尽量不跳步	数理步骤跳跃， 结论先行
素养 维度	情境真实自洽，能体现数学应用价值	物理尺度失真， 跨学科情境流于表面

（二）人机协同命题的“双漏斗”机制

随着生成式 AI 介入命题过程，传统上由教师独立完成的命题流程，逐渐转变为一种“先扩展、再筛选、后校正”的协同模式。结合本研究的修订记录来看，这一过程大体可归纳为五个环节：首先明确考点与命题目标，其次将需求转化为提示词，再由模型生成初稿，随即由教师从科学性 with 教学适切性两个维度进行筛查，最后经过专家核查与版式整理形成定稿。整体流程并不繁复，但其中的人工审校环节，往往直接决定试题能否真正进入课堂或试卷。

三、教师问卷的实证分析

（一）样本特征与使用概况

本研究样本主要来自教学要求较高的高中，相关教师对试题命制的科学性与严密性通常具有较强的辨识能力。192 份有效问卷中，市级重点中学教师占 53.6%（103 人），区级重点及其他学校教师占 38.0%（73 人），其余为 8.4%（16 人）。从教龄分布看，10 年以上教龄教师占 66.1%（127 人），4 至 10 年教龄者占 23.4%（45 人），3 年以下者占 10.5%（20 人）。整体而言，样本具有较好的层次性与代表性。

（二）现实障碍

李克特五点量表的统计结果显示，教师对 AI 参与命题的总体接受度处于中上水平，均值为 3.41 分。多数教师认可其在初稿生成与版式处理方面的

表 2 教师对生成式 AI 辅助命题的多维评价
(样本量 n=192)

评价维度	均值 (5分制)	标准差	维度解释
总体满意度	3.41	0.68	整体评价处于中上水平, 仍有改进空间
效率提升度	3.52	0.63	初稿构思和排版时间明显缩短
情境创新性	3.41	0.71	跨学科素材较新, 新鲜感较强
知识综合性	3.45	0.70	多知识模块交叉能力较好
数理严谨性	3.26	0.78	得分最低, 仍依赖人工深度审校

效率优势, 但同时也指出了较为明显的使用门槛: 70.98% 的受访者认为高质量提示词并不容易把握; 超过半数教师则指出 AI 在知识覆盖的准确性与数理推演的严谨性方面仍存在不足。由此可见, 模型可以较快完成“形式上的生成”, 却难以直接达到专业命题所要求的“实质性贴合”。

(三) 经验分化与干预深度的关联

变量交叉检验表明, 教龄与教师对 AI 功能的判断存在显著相关性 ($\chi^2=14.23$, $df=4$, $p<0.01$)。质性资料进一步显示, 教龄越长的教师, 越重视学科边界与试题规范, 因此也更倾向于将 AI 定位为素材生成工具, 而非可以替代全流程的命题主体。相应地, 其审核重点也由“是否成题”逐渐转向题目结构是否完整、推理链条是否闭合, 以及情境设置是否真实可信。

(四) 开放式质性反馈挖掘

开放式反馈进一步印证了上述判断。有教师指出, AI 生成的解析几何题常因参数设置不当而失真; 也有教师认为, 所谓“新定义”试题往往语义繁复、指向不清。总体看, AI 的主要问题不在排版, 而在隐含约束、数理逻辑与边界条件上。因此, 人机协同实质上是对机器初稿进行持续勘误与重构。

四、典型案例的人机协同迭代分析

以下案例采用“问题背景—初稿问题—教师处理—定稿呈现—分析说明”的逻辑结构, 保留修订过程, 同时呈现试题从初稿到定稿的调整路径。

(一) 解析几何：情境尺度与参数常识的校准

问题背景：本题以双曲线标准方程和离心率为核心, 重点考查参数关系与情境设置是否协调。

AI 初稿缺陷：初稿对“玉壶春瓶”的尺度设定偏大, 虽然题目结构完整, 但与日常陈设的常识不符, 情境显得不够贴切。

教师干预策略：修订时, 教师没有继续扩展背景, 而是将器物尺度调整到案头陈设的合理范围内, 并重新核对瓶身轮廓、焦点距离与实轴长之间的对应关系, 使题目回到可理解、可计算的轨道上。

提示词关键调整：“设计一道用双曲线建模玉壶春瓶的选择题。器物高度应符合案头陈设常识, 约 30cm; 瓶颈最细处外径对应双曲线实轴长; 参数设置须满足离心率公式, 且不得出现退化情形。”

定稿呈现：（单选题）我国古代“玉壶春瓶”其瓶身轮廓可近似看作双曲线。已知双曲线中心在原点, 瓶颈最细处外径（实轴长）为 8cm。若瓶身上一点 P 在右支上, 到左焦点距离为 14cm, 到 $x=\frac{a^2}{c}$ 距离为 4cm, 则离心率为（ ）。

- A. 5/4
- B. 3/2
- C. $\sqrt{2}$
- D. 2

（正确答案：B）

深度分析：情境一旦失真, 应用题就很难真正成立。教师的修订, 实际上是把题目从“有叙述”升华为“情境与参数相互匹配”。

(二) 函数与代数：离散对象与连续方法的界限

问题背景：本题考查递推数列的单调性与不等式放缩, 关键在于推理链条是否闭合。

AI 初稿缺陷：初稿在处理放缩证明时, 直接调

用连续函数的定积分方法,试图解决离散数列问题。表面上看,这样处理比较顺畅,但实际上混淆了研究对象,把定义在正整数集上的离散变量当成连续函数进行处理。

教师干预策略:教师发现这一问题后,及时切断了模型对连续分析工具的误用,要求其先从代数角度判断区间内的有界性,再依据递推关系逐项展开,回到离散推理本身。

定稿呈现:(多选题)已知正项数列 $\{a_n\}$ 满足 $a_1 \in (0,1)$, 且 $a_{n+1}=a_n-a_n^2 (n \in \mathbb{N}^*)$ 。下列论断正确的有()。

- A. 单调递减且有下界
- B. $a_n < \frac{1}{n}$
- C. $a_n < \frac{1}{n+1}$
- D. 前 n 项平方和小于 1

(正确答案: ABD)

深度分析: 离散对象和连续对象在数学底层结构上完全是两码事。要是把积分等连续工具生搬硬套到高中数列里, 不仅削弱了推理的严密性, 还容易误导学生, 让他们养成机械套公式的坏习惯。

(三) 立体几何: 空间指代与几何关系的精准表述

问题背景: 本题将长方体作为研究对象, 侧重考查空间里各种点线面位置关系以及概念定义。

AI 初稿缺陷: AI 生成的“定体量”新概念虽具有一定新颖性, 但在描述异面直线距离等空间属性时, 语言冗余、修饰繁杂, 致使题意表达含混不清。学生只读题干, 难以在头脑中构建出稳定清晰的空

间几何画面。

教师干预策略: 针对 AI 初稿中表述含混的问题, 教师采取精简优化策略, 直接剔除冗余修饰与易引发歧义的描述, 代之以底面周长、侧面积、体对角线等具体可量化的几何指标。通过减少信息干扰、突出核心变量, 题目指向性与可操作性均得到显著提升, 能有效帮助学生建立清晰的空间几何认知框架。

定稿呈现:(单选题)在长方体中,若一组几何量

能唯一确定其体积,则称这组量为“定体量”。下列四组几何量中不能成为长方体“定体量”的是()。

- A. 互不相邻的三个面的面积
- B. 两个相邻侧面的面积
- C. 从同一顶点出发的三条对角线长
- D. 从同一顶点出发的三条棱长

(正确答案: B)

深度分析: 日常生活语言往往具有一定的灵活性和模糊性,但数学语言必须追求精确性与唯一性。用明确的几何量替代含混表述, 不仅能减少审题歧义, 也更符合立体几何对空间关系精确表达的要求。

(四) 概率统计: 样本空间边界与底层规则的重塑

问题背景: 该题借用二进制编码, 考查离散型随机变量的分布列和数学期望, 解题关键精准界定样本空间。

AI 初稿缺陷: AI 在底层设定上出现隐性失误: 在初始化全集时, 系统默认二进制的最高位可以选

表 3 四类典型试题的人机协同修订轨迹与干预靶点

模块归属	AI 初稿核心缺陷	修订轮次	教师干预提示词关键调整方向	定稿终局状态
解析几何	生活常识脱节, 物理尺寸失真	3 轮	补充限制并对双曲线参数约束	情境与参数重新对齐
函数代数	忽略离散性, 误入连续化	3 轮	禁止积分放缩, 回归逐项推导	推理闭环, 回到离散逻辑
立体几何	空间指代含糊	2 轮	删除模糊表述, 改用可量化几何量	几何对象唯一, 图形更清楚
概率统计	前提约束缺失	2 轮	锁定最高位为 1, 重建基本事件空间	样本空间成立, 模型可解

0。这种常识性失误直接导致样本空间界定失准，后续算概率、算期望全跟着错了。

教师干预策略：教师介入时，首先要严格守住样本空间的理论边界，在修改时，特意在提示词里强调了“二进制数首位必须是1”这条规则。该前提一经确立，后续随机变量讨论和数理计算才立得住。

定稿呈现：（综合题）设全体 n 位二进制数构成集合 S ($n \geq 3$)，且最高位固定为1。从 S 中随机抽取一个二进制数，记其中0的个数为 X 。（1）当 $n=5$ 时，求概率 $P(X=2)$ ；（2）设 $P(X=k)$ 最大时对应的 k 值为 m ，求 m 的表达式。

深度分析：样本空间一旦界定有误，后续计算就没有意义了。教师在此环节的作用，在于首先重建概率推理所依赖的逻辑前提。

（五）跨案例综合透视

综合四类题目的修订过程可以看出，AI更擅长题材发散、素材组织和框架铺陈；而教师的核心价值，则集中体现在边界约束、逻辑补缝和生活常识校准。虽然人工介入的路径各不相同，但教师干预命题的目标是一致的：让模型给出的可能越界的内容回归高中数学评价所要求的严密、规范与可判定之中。

五、讨论与结论

（一）研究成效透析

从修订日志看，人机协同确实提高了命题前期效率：初稿生成时间约缩短45%，整体工时约下降

30%。不过，这种效率优势主要体现在起草阶段；一旦进入逻辑校验、边界修补和语言定稿，教师投入的时间与精力并没有明显减少。也就是说，AI提升的是初始效率，后续对命题质量的把关仍需要依赖人工判断。

（二）对数学教育评价的启示

本研究说明，自动化工具正在改变数学试题评价的关注重点^[4]。知识层面，教师不能只看学生对公式记得牢不牢，更要看条件识别是否准确；能力层面，应更重视学生对离散与连续的区分、对象与方法的匹配等关键问题；素养层面，不能停留在简单“贴标签”，而要保证情境描述、数学逻辑与现实常识彼此一致。

（三）研究局限及展望

本研究样本主要来自重点高中，因此结论的推广范围仍需进一步验证。与此同时，大语言模型更新迭代较快，同一提示词在不同版本中的效果也可能存在差异。后续研究可进一步扩大地区样本，并比较不同模型版本下的人机协同修订表现。

（四）结论

生成式AI并没有削弱高中数学命题的专业门槛，反而把教师对题目边界、逻辑和表达的把关要求抬得更高。面对容易在边界处出错的大模型，教师真正需要关注的，不再是机械重复的出题动作，而是发现问题、修正问题和判断问题是否成立的能力。想要人机协同走得更远，可以把发散性的资料搜索交给机器，但必须将严密的数学判断留给专业教师。

参考文献

- [1] 曹一鸣, 吴景峰. 生成式AI赋能数学课堂教学内容选配的探索与研究: 以高中数学例题习题选配为例[N]. 数学教育学报, 2024, 33(5): 60-66.
- [2] 吴河江, 吴砥. 生成式人工智能教育应用: 发展历史、国际态势与未来展望[J]. 比较教育研究, 2024, 46(6): 13-23.
- [3] 喻平. 数学核心素养评价的一个框架[J]. 数学教育学报, 2017, 26(2): 19-23+59.
- [4] 任子朝, 陈昂, 赵轩. 数学核心素养评价研究[J]. 课程·教材·教法, 2018, 38(5): 116-121.