

初高中数学错题智能体技能的设计与实践

郑 富^{1*}, 王雨墨²

¹河南大学生命教育中心, 河南 开封

²石家庄第四十二中学, 河北 石家庄

收稿日期: 2026年4月6日; 录用日期: 2026年5月8日; 发布日期: 2026年5月14日

摘 要

本研究面向初高中数学错题处理场景, 提出一套基于Agent Skill架构的错题智能体设计框架。该框架围绕“采集-解析-诊断-报告”闭环 workflow, 将题目解析、错因诊断、学情分析与对话辅导封装为四个可调用的技能模块, 并构建覆盖概念性、程序步骤性、计算性、符号表示性四类错误的错因模板库, 将诊断输出与课程标准对齐, 实现从“知道错了”到“知道怎么改”的可执行反馈。研究同时提出省时、高效、针对性三维评估指标体系, 并就AI辅助学习的伦理边界展开讨论。

关键词

错题智能体, 形成性评价, 错因诊断, Agent Skill, 迁移学习, 数学教育

Design and Practice of Intelligent Agent Skills for Wrong-Answer Analysis in Junior and Senior High School Mathematics

Fu Zheng^{1*}, Yumo Wang²

¹Life Education Center, Henan University, Kaifeng Henan

²The 42nd Middle School of Shijiazhuang, Shijiazhuang Hebei

Received: April 6, 2026; accepted: May 8, 2026; published: May 14, 2026

Abstract

This study addresses the problem of handling wrong answers in junior and senior high school mathematics, proposing a design framework for a wrong-answer intelligent agent based on the Agent

*第一作者。

文章引用: 郑富, 王雨墨. 初高中数学错题智能体技能的设计与实践[J]. 教育进展, 2026, 16(5): 774-782.

DOI: 10.12677/ae.2026.165920

Skill architecture. Centered on a closed-loop workflow of “collection-parsing-diagnosis-reporting”, the framework encapsulates four callable skill modules—question parsing, error diagnosis, learning situation analysis, and conversational tutoring—and constructs an error template library covering four error types: conceptual, procedural, computational, and symbolic. Diagnostic outputs are aligned with curriculum standards to deliver executable feedback that moves students from “knowing what went wrong” to “knowing how to improve”. The study further proposes a three-dimensional evaluation framework encompassing time efficiency, learning gains, and individual differentiation, and discusses the ethical boundaries of AI-assisted learning.

Keywords

Wrong-Answer Intelligent Agent, Formative Assessment, Error Diagnosis, Agent Skill, Transfer Learning, Mathematics Education

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

20 世纪以来, 学习科学、认知心理学与教育评价研究的持续推进, 深刻重构了研究者对“人类学习”的基本理解。早期研究聚焦教学行为本身, 追问“教师如何施教, 才能有效传递知识与技能”; 随着研究重心从“有效传递”转向“有效学习”, 学界开始强调学生如何在任务与证据中调整理解、建构结构化知识并实现迁移。这一转向使课堂中的核心问题发生根本性改变: 不仅要回答“教了什么”, 更要回答“学生如何学会、为何学不会、怎样改进”。

然而, 这一理论转向在实际课堂中的落地仍面临明显缺口。在初高中数学情境下, 数学错误往往并非偶发失误, 而是概念边界不清、推理链条存在缺口、方法选择不当或表征转换受阻等结构性问题的外显。现有错题处理方式——以“错题本”为代表——维护成本高、容易任务化, 多停留于单题复现与“做类题”, 对错因的系统解释与可操作改进路径支持严重不足。

为理解这一落差, 需回到现代学习环境的两条基本原理。其一, 有效的学习环境须伴有明确的学习期待, 并能展开形成性评价——反馈的价值在于让学生将“结果偏差”转化为“可解释的差距”并获得可执行的改进路径。其二, 有效的学习成效体现在学生能否将知识组织为可调度的层级结构并在新情境中迁移运用。这两条原理在过去三十余年的课堂评价与认知科学研究中已经积累了大量实证支撑, 被视为设计有效学习环境的基本前提。生成式 AI 与智能体技术的普及, 为上述原理在数学错题处理中的落实提供了新的实现条件。

鉴于此, 本研究聚焦四个核心问题: 其一, 错题智能体应具备哪些关键技能, 如何进行模块化设计以支撑“采集-诊断-报告”的闭环; 其二, 如何实现面向初高中数学的错因诊断与可解释表达, 使反馈真正对齐学习期待; 其三, 错题智能体如何服务教师的纠错教学与学生的自主修正; 其四, 如何从学习结果与学习过程两条路径评估其实践效应。

2. 相关研究与理论基础

2.1. 学习科学视角下的错误价值与形成性评价

在错误的认知地位问题上, 发生认识论与社会文化学习理论的奠基性研究早已确立了一个基本共识:

错误并非学习的对立面, 而是知识建构过程中不可避免的自然产物, 其实质是学生当前认知图式与目标知识结构之间张力的外部显现。后续认知诊断领域的研究沿着这一思路进一步拓展, 揭示了错误在评价功能上的独特价值——通过对错误模式的系统分析, 研究者可以推断学生在特定知识属性上的掌握状态, 进而为精准的教学干预提供认知层面的依据。换言之, 错误本身携带着高密度的诊断信息, 关键不在于消除它, 而在于如何将其系统化地提取并转化为教学行动的起点。

在反馈与形成性评价的研究脉络中, 学界经过数十年的积累已形成一组核心判断。首先, 有效反馈的功能重心不在于告知对错本身, 而在于帮助学生清晰地理解“自己当前处于什么位置”与“目标要求是什么”之间的落差, 以及这一落差可以通过何种路径被弥合。其次, 形成性评价要真正发挥促进学习的作用, 须同时满足三个前提条件——学生对目标标准有明确认知、具备监控自身表现的意识与手段、能够将所发现的差距转化为可操作的改进行动。三者缺一, 反馈便停留在信息传递的层面而无法内化为学习调节机制。此外, 认知科学关于专家与新手差异的长期研究反复印证了一条规律: 能否将零散知识组织为具有层级结构的可调度体系, 是决定知识能否跨情境迁移的关键前提。这一发现直接为本研究的错因诊断逻辑与结构化知识封装设计提供了理论锚点。

2.2. 数学错误的认知分析与诊断研究

在错误分类体系的构建上, 数学教育研究领域经过近半个世纪的理论探索与实证积累, 已经发展出较为成熟的多维归因框架。早期具有开创意义的研究从认知加工的视角出发, 将数学错误的来源归纳为语义编码障碍、概念缺陷、规则应用不当、策略选择失误与负迁移干扰等若干范畴, 这一分类思路至今仍是后续研究的重要参照基线。在代数学习领域, 大量课堂观察与临床访谈研究发现, 学生对等号的理解长期停留在“操作性”层面——即把等号视为“算出答案”的指令信号, 而非两个数学表达式之间等价关系的声明——这种深层的概念偏差被反复证实是相当一部分代数错误的隐性根源。在函数学习领域, 一个被广泛讨论的现象是学生头脑中所构建的“概念映像”与数学上严格的“概念定义”之间存在持续性的冲突与张力: 比如初中里有些同学觉得“函数就是一条连续的曲线、一个公式 $y = \dots$ ”(概念映像), 但严格定义是“每个 x 只能对应一个 y ”(概念定义)。因此遇到分段函数、图像有断点或空点时, 学生会凭直觉误判“这不是函数”或把定义域、值域写错, 导致反复出错。国内相关研究方面, 周蓓与尹绍军指出, 高中数学学习中的“知识误解”现象普遍存在, “概念性误解是指学生在学习数学概念时, 由于理解不准确或缺乏深入思考, 导致对数学概念的定义和内涵产生偏差”, 且此类误解“比较隐蔽”, 难以及时察觉与干预[1]。在运算错误的类型分布上, 王思宇等发现, “七年级学生在运算中出现的错误, 总体来说占比最大的就是概念性的错误”, 并进一步区分了概念性、程序性、策略性与疏忽性四类错误成因, 指出各类错误在不同教学情境下具有不同的诊断与干预价值[2]。在错题分析的实践困境上, 高陈欣玥观察到, “大部分学生只是在原题上修改, 或者是抄题后在本上修改, 但极少有同学会真正地分析自己的错因”, 学生在长期整理中“容易把错题分析看作惩罚”, 导致错题改正率持续偏低[3]; 在教师层面, 该研究同样指出, 目前错题分析研究“尚未出现关于不同错题分析方法的定位、顺序及整合的系统性研究”, 教师缺乏可操作的整合工具以系统识别全班共性问题[3]。

2.3. 生成式 AI 与智能体技术的教育应用进展

近两年来, 随着大语言模型能力边界的快速拓展, 其在数学教育相关任务上的表现已引起广泛关注。多项独立评估表明, 当前主流大模型在数学题目语义解析、解题步骤生成及常见错误模式识别等任务上已展现出相当可观的能力水平, 部分结构化推理任务上的表现甚至接近经过专门训练的人类评判者。但同样不容忽视的是, 大模型固有的“幻觉”倾向——即在缺乏充分外部约束的情况下, 生成看似自洽实

则事实性错误的输出——对教育应用场景构成了不可低估的可靠性风险。在面向学生的错因诊断这一高风险任务中, 一条错误的诊断结论可能误导学生的后续学习方向, 因此在系统设计层面建立严格的生成约束与人工可校正机制, 不是锦上添花的选项, 而是底线性的要求。

在智能体架构的设计范式上, 近年来的工程实践与研究探索已经表明, 将复杂任务拆解为多个独立可调用的技能模块(Skill), 是在保持系统整体灵活性的同时提升单项任务执行稳定性与结果可追溯性的有效策略。这一思路与软件工程中模块化、高内聚低耦合的经典设计原则一脉相承, 其在教育场景中的优势在于: 每个技能模块的输入输出边界清晰可控, 便于教师理解系统的决策依据, 也便于在出现误判时定位问题环节并进行针对性修正。与此同时, 教育技术领域的主流共识日益明确: AI在教育评价中的恰当定位是“辅助决策者”而非“替代决策者”——它应当增强教师获取诊断信息的效率与深度, 而非绕过教师的专业判断直接对学生施加干预。

综上可见, 现有研究在“形成性评价闭环+知识结构化+迁移支持”三者同时具备的层面上仍存在显著空白, 本研究由此切入。

3. 错题智能体技能的设计原则与总体架构

“错题智能体”是基于 Agent Skill 架构设计的智能教育工具, 能够自动采集学生作答中的错误, 进行结构化解析与类型诊断, 并生成面向教师和学生的差异化反馈。与传统错题本不同, 错题智能体将整个处理流程——从采集、解析、诊断到报告——封装为可调用的技能模块(Skill), 以诊断为驱动核心, 实现从“记录错题”到“诊断错因、提供可执行反馈”的功能升级。

3.1. 设计目标与原则

本研究的系统设计基于三条核心原则, 每条均与形成性评价理论及评估维度直接对应。

首先是以诊断为核心。系统的首要目标是“定位错误根源”, 而非简单“记录错题”。每个功能模块都服务于错因的识别、解释与传达, 系统输出必须给出具体的错误类型标注与知识点定位, 不得停留在“该题答案有误”的层面。

其次是数据驱动与可解释性统一。诊断逻辑基于结构化知识体系和错误分类规则, 而非依赖大模型的黑盒输出。每条诊断结论须附有可追溯的证据(对应知识点、错误类型及判断依据), 使教师能够核查与修正。

最后以三维指标锚定设计决策。每项功能设计须对应以下三个评估维度之一: 时间成本(是否有助于降低错题处理时间)、学习收益(是否有助于提升知识点掌握或降低重复错误率)、个体差异适配(是否能根据学生错误特征提供差异化反馈)。

3.2. 功能模块与技能清单

基于上述原则, 本研究将错题智能体的核心能力按 workflow 分解为四个功能模块。

首先是采集,指将输入题目(图片、PDF 或文字格式)转化为结构化表示, 并自动标注知识点、题型与能力要求层级, 是后续诊断功能的数据基础。

其次是解析,指结合题目解析结果与学生作答, 识别错误所属类型(概念性/程序步骤性/计算性/符号表示性), 调用错因模板库生成具体的错因描述与知识点定位。

再次是诊断,指在班级层面聚合个体诊断数据, 生成面向教师的学情报告, 包括班级错因热力图、高频错因聚类、分层作业建议及讲评要点提取。

最后是分析报告,指为学生提供基于错因诊断的智能体对话支持, 允许学生就诊断结论追问“为什

么”、就纠错步骤追问“怎么做”，并动态调整解释的详略程度。

四个功能模块共同构成“采集-解析-诊断-报告”的闭环 workflow，各模块均以独立的 Agent Skill 形式封装，具备标准化的输入输出接口，可在不同教学场景下灵活组合调用。

3.3. 质量保障机制：可解释性、可校正性与可追溯性

首先是可解释性约束，所有诊断输出须遵循“结论 + 证据 + 知识点定位”的结构，每条错因诊断必须包含错误类型标签、触发诊断的具体错误特征及知识点编码，将大模型的生成空间限定在有据可查的范围内。

其次是可校正性设计，系统为教师提供对诊断结论的修正功能，可直接修改错因类型标签和证据描述，修正后结果写回错因模板库，形成人机协作的持续迭代闭环。

最后是可追溯性架构，每条系统输出保留其生成所依据的规则来源与知识点引用，支持事后的版本回溯，既是质量管理的基础，也是系统迭代优化的数据来源。

4. 错因诊断与知识结构

4.1. 大语言模型驱动方式与 Agent Skill 运行机制

本系统以 DeepSeek 671B 大语言模型作为底层推理引擎，并通过 API 接口方式接入 Agent Skill 架构，实现错题解析、错因诊断与反馈生成等核心功能。与将大模型直接作为“答案生成器”的做法不同，本研究将其定位为受约束的推理组件：模型主要负责对题目文本、学生作答文本和结构化知识约束进行综合理解，并在预设流程中输出标准化诊断结果，而非脱离规则自由生成结论。

在运行机制上，系统由四个相互衔接的层级构成。第一层为**输入处理层**，负责接收图片、PDF、Word 和文本等不同格式的题目与作答数据，并完成识别、清洗和字段标准化；第二层为**模型推理层**，调用 DeepSeek 671B 模型，对题目语义、作答过程和错误特征进行分析；第三层为**知识约束层**，由课程标准知识点编码、错误类型判定规则和分年级数学错因 JSON 数据集构成，用于约束模型推理边界；第四层为**结果输出层**，将模型结果整理为结构化诊断 JSON、教师端学情摘要和学生端可理解反馈。

该技术路线的核心不在于单纯依赖大模型能力，而在于通过“模型推理 + 规则约束 + 数据模板”的协同机制，提高诊断结果的可解释性、可校正性与可追溯性。具体而言，DeepSeek 671B 模型负责完成语义理解和推理生成，结构化提示词负责限定任务流程和输出格式，分年级数学错因 JSON 数据集则提供领域知识与模板支撑，从而降低纯生成式输出中可能出现的幻觉和随意归因问题。采用 API 接入方式而非本地端到端部署，除了便于在不同教学平台和错题处理场景中快速调用统一模型能力，也使系统能够将提示词、知识模板和输出格式控制集中在服务端管理，从而提高版本迭代效率与诊断一致性。

4.2. 结构化提示词设计

为使大语言模型在错因诊断任务中保持稳定、可控和可复核的输出，本研究采用结构化提示词设计方式，对模型的角色定位、推理流程和输出格式进行分层约束。整体提示词由三个部分组成：角色定义层、诊断流程层与输出约束层。

角色定义层用于明确模型任务边界，例如要求模型扮演“初高中数学错因诊断助手”，任务目标不是直接给出题目答案，而是依据题目、学生作答和错误规则完成错误类型判定与诊断解释。诊断流程层则以分步指令方式，要求模型依次执行“定位错误位置 - 匹配知识点 - 对照错误类型判定规则 - 调用对应错因模板 - 输出诊断结果”的流程。输出约束层要求模型严格按照预设 JSON Schema 返回结果，不得生成额外自由文本。

本研究采用的错因诊断提示词(节选)

角色: 你是一名初高中数学错因诊断助手。你的任务不是直接给出标准答案, 而是根据题目、学生作答、知识点规则和错误类型标准, 对学生错误进行分类并给出解释。

输入: 年级: 高一

题目: 解不等式 $x^2 - 5x + 6 > 0$

学生作答: $x > 3$ 或 $x > 2$ 即 $x > 2$

知识点候选: 一元二次不等式; 因式分解; 解集表达

流程要求:

1) 定位出错步骤; 2) 判断主要错误类型; 3) 引用知识点编码; 4) 生成错因解释与纠错建议; 5) 按照 JSON 输出。

输出字段: error_type、knowledge_code、evidence、diagnosis、correction_strategy、confidence

通过上述结构化提示方式, 模型的输出空间被限制在“有规则、有模板、有格式”的范围内, 使其更适合教育评价场景中对稳定性和可解释性的要求。

4.3. 分年级数学错因 JSON 数据集的构建

为提高错因诊断的年级适配性和知识定位精度, 本研究构建了分年级数学错因 JSON 数据集, 并将其作为 Agent Skill 的外部知识支撑。数据集按初中七至九年级和高中一至三年级分别建立子库, 每个子库内部按照“学段 - 主题 - 知识点 - 错误类型 - 错因模板”的层级组织。

数据集主要来自三个方面: 第一, 初高中数学试卷、练习题和作业中的典型错题样本; 第二, 一线教师在作业批改、课堂讲评和错题整理中总结的高频错误表述; 第三, 研究团队依据课程标准、教材目录和教学经验提炼形成的知识点编码体系与错因描述模板。通过多来源整合, 尽可能避免数据来源单一所带来的结构性偏差。数据集形成主要经历四个步骤。第一, 对学生错题样本进行收集与整理, 提取题目、学生作答及错误片段; 第二, 依据课程标准和教材章节对题目进行知识点编码; 第三, 由研究者与教师共同依据四类错误类型框架完成错因归类; 第四, 将相近的错误实例抽象为统一模板, 并补充相应的纠错建议和反馈文本。

在质量控制方面, 本文采用“研究者初标 - 教师复核 - 分歧协商”的方式提高数据可靠性。对于归类不一致的样本, 不直接进入模板库, 而是在协商一致后再纳入正式数据集。与此同时, 系统在实际运行中保留教师对诊断结果的修正记录, 并将高频修正项回写到 JSON 模板库中, 用于后续迭代更新。在大模型使用中, 选用的 DeepSeek 671B 模型并不直接记忆所有数学错因, 而是通过 API 调用时读取与当前年级、知识点相匹配的 JSON 子库, 将相关模板作为上下文约束输入模型, 从而实现“模型负责推理, 数据负责约束”的协同诊断机制。

4.4. 错误类型与判定规则

Table 1. Error type ontology: definitions, determination features, and structural problem mapping

表 1. 错误类型本体: 定义、判定特征与结构性问题对照

错误类型	形式化定义	典型判定特征	对应结构性问题
概念性错误	学生对某一数学概念的定义、边界条件或适用范围存在错误认知, 导致在概念识别或应用阶段出错	定义混用; 边界判断错误; 概念外延扩大或缩小	概念边界不清; 条件意识薄弱
程序步骤性错误	学生理解题目意图, 但在执行解题步骤的序列、条件判断或分支处理时出现遗漏、颠倒或错误跳转	步骤遗漏; 条件判断缺失; 算法路径选择错误	推理链条缺口; 方法选择不当

续表

计算性错误	解题思路与步骤框架正确,但在具体数值运算、代数化简或方程求解的执行层面出现错误	运算符号错误;化简步骤错误;数值计算失误	推理链条中的执行层缺口
符号表示性错误	学生在数学符号的书写、使用或解读层面出现错误,包括符号混淆、表示规范不符或符号语义理解偏差	符号混淆;书写不规范;符号语义误读	表征转换受阻

在多类错误并发时,系统按“概念性 > 程序步骤性 > 计算性 > 符号表示性”的优先级输出主要错因,并在次要字段记录其他错误特征(见表 1)。此外,系统设置“粗心失误”类别(优先级最低),并在输出该标签时附加提示:建议教师核查是否存在被遮掩的概念性理解问题,避免将深层错误简单归因为粗心。

5. 实践实施: workflow 设计

5.1 workflow 闭环: 采集 - 解析 - 诊断 - 报告

第一环节: 采集。系统支持三类输入格式:纸质试卷图片(手机拍照或扫描仪)、电子文档(PDF 或 Word 格式)、平台导出数据(与作业平台对接的结构化数据)。图片输入调用多模态识别能力提取题目文字与数学公式;电子文档直接进行结构化解析。输出为标准化的“题目 - 作答”数据对,含原始题目文本、学生作答文本及采集时间与来源标注。

第二环节: 解析。系统对采集数据进行知识点标注,调用知识点查询 Skill 将题目映射到结构化知识体系,标注题型与能力要求层级。输出为“题目信息卡”,包含题目文本、知识点编码列表、题型标签、能力要求层级及置信度评分(标记需人工复核的低置信度标注)。

第三环节: 诊断。系统基于“题目信息卡”与学生作答,调用错误类型识别逻辑与错因模板库,生成诊断报告。输出包含错误类型标签、错因描述(含证据)、知识点编码及能力要求层级映射。低置信度诊断结论自动标记并推送至教师端进行人工核查。

第四环节: 报告。系统在班级层面聚合个体诊断数据,生成可视化学情报告,包含班级错因热力图、高频错因列表、按错因分组的学生名单、分层作业建议及讲评要点摘要,支持导出为 PDF 或表格格式。

5.2. 典型范例

本节以一个典型案例验证系统的诊断逻辑,对应程序步骤性错误这一核心错误类型,按“输入→诊断输出→教师端信息”的统一格式呈现。

案例: 程序步骤性错误——解一元二次不等式

【输入】题目: 解不等式 $x^2 - 5x + 6 > 0$ 。学生作答: $x > 3$ 或 $x > 2$, 即 $x > 2$ 。

【诊断输出】错误类型: 程序步骤性错误。错因描述: 学生完成了因式分解 $(x-2)(x-3) > 0$, 但在由因式分解结果推导解集时,将“乘积大于零”等同于“两个因子都大于零”,忽略了“两个因子都小于零”的情况。知识点定位: 一元二次不等式 - 解集推导的完备性处理。

【反馈文本 - 任务层】你的因式分解 $(x-2)(x-3) > 0$ 是正确的。但乘积大于零有两种情况: (1) 两个因子都大于零: $x > 2$ 且 $x > 3$, 即 $x > 3$; (2) 两个因子都小于零: $x < 2$ 且 $x < 3$, 即 $x < 2$ 。两种情况取并集,解集为 $x < 2$ 或 $x > 3$ 。

【反馈文本 - 过程层】你目前的解题流程在“因式分解后如何讨论符号”这个步骤上有遗漏。建议建立固定的分情况讨论框架: 乘积 > 0 , 列两种情况(同正/同负); 乘积 < 0 , 列两种情况(一正一负)。

【反馈文本 - 自我调节层】以后遇到乘积类不等式时,可以在动手前先问自己: 乘积为正有几种情

况? 乘积为负有几种情况? 确认已列出所有情况再求解集。

【教师端信息】本班在此类问题中共有 8 名学生出现“仅讨论同正情况”的错误, 6 名学生出现解集表达格式错误(如用“且”代替“或”)。建议分层处理: 前 8 名学生需专项补充“乘积正负的完备分情况”练习; 后 6 名学生重点强化解集表达规范。

6. 效果评估方案

评估指标体系

本研究构建了三个维度的评估指标体系, 每个维度包含可量化的客观指标与可测量的主观指标。

第一维度: 省时(时间成本指标)。客观指标包括: 学生错题整理时间(目标: 较基线缩短 20% 以上)、教师批注时间(系统使用前后对比, 目标: 单次作业批阅时间减少 15 分钟以上)、教师讲评备课时间(目标: 从 45 分钟缩短至 20 分钟以内)。主观指标为教师对系统节省工作负担程度的 5 分制评分(目标均值 ≥ 4 分)。

第二维度: 高效(学习收益指标)。客观指标包括: 重复错误率(以 3 次以上测评的纵向数据计算, 目标: 错误率下降 30% 以上)、知识点掌握曲线(对系统识别出存在错因的知识点, 后续测评正确率变化)、迁移题正确率(针对性反馈后在相关迁移题上的表现, 目标: 正确率提升 15% 以上)。主观指标为学生对系统学习帮助度的 5 分制评分(目标均值 ≥ 3.8 分)。

第三维度: 针对性(个体差异适配指标)。客观指标包括: 建议采纳率(教师采纳系统分层作业建议的比例, 目标 $\geq 70\%$)、诊断修正率(教师对系统诊断结论进行修正的比例, 目标修正率 $< 20\%$)。主观指标为教师对系统针对性反馈质量的 5 分制评分(目标均值 ≥ 4 分)。

7. 讨论

本研究提出的错题智能体设计框架并非普适于所有数学题型, 其有效性在不同题型与场景下存在明显的边界差异。系统在以下情况下表现最为稳定: 知识点边界清晰的题目(如代数运算、方程求解、基础函数题)、错误类型属于本研究定义的四类主要类型的题目, 以及错因模板库中已有对应覆盖的题目。

相对而言, 系统在开放性推理题(如几何证明、数学归纳法论证)、跨章节综合题以及学生作答极不规范的情况下可靠性有所下降。对于上述情况, 系统应自动标记为“需人工诊断”推送至教师端, 而非强行输出低置信度的诊断结论。

在设计与实践贡献层面, 本研究提出的“采集-解析-诊断-报告”四环节闭环 workflow, 以及基于 Agent Skill 的知识体系封装方式, 为在资源约束条件下实现数学纠错的智能化提供了切实可行的技术路径。在未来的后续研究中, 可在以下方向深入探索: 一是进一步优化系统的个性化反馈机制; 二是扩展至其他理科学科, 验证框架的跨学科迁移性; 三是在更大规模的准实验研究中系统评估长期学习效果, 并深入研究 AI 辅助学习的伦理边界与数据保护机制。

需要强调的是, 本研究将 AI 定位为辅助教师决策的工具, 而非替代教师专业判断的自动化系统。在 AI 辅助学习深度嵌入教育实践的当下, 坚持教师监督权、学生数据保护与诊断可解释性这三项原则, 不仅是技术设计的选择, 更是教育伦理的底线。

8. 伦理、风险与研究局限

8.1. 自动化偏见风险

尽管本系统通过结构化提示词、错误类型判定规则与分年级数学错因 JSON 数据集对模型输出进行约束, 但自动化诊断仍可能存在偏见风险。一方面, 若模板和样本更多覆盖常见题型、规范表达或标准

化解法, 系统可能对非常规作答、跳步表达或特殊解法识别不足; 另一方面, 多类错误并发时的优先级规则虽然有助于提高输出一致性, 但也可能压缩学生真实的复杂错误结构。为降低上述风险, 系统对低置信度结果和开放性题型保留教师复核机制, 并将教师修正结果回写模板库, 以持续减少系统性偏差。

8.2. 对教学实践的潜在负面影响

本系统能够提升错题分析效率, 但若使用不当, 也可能对教学实践造成负面影响。教师若过度依赖系统诊断, 可能削弱对学生真实思维过程的深入追问; 学生若直接接受系统生成的错因解释, 可能减少自主反思和自我修正过程; 系统更适合边界清晰、结构明确的题型, 可能在使用中强化对“易于自动分析内容”的关注。基于上述影响, 本研究坚持将 AI 定位为辅助教师决策的工具, 而非替代教师专业判断的自动化系统。

8.3. 数据管理与隐私保护框架

针对题目文本、学生作答和诊断结果等学习数据, 以“最小化采集 - 脱敏处理 - 分级授权 - 留痕审计”的数据管理框架, 仅采集完成错因诊断所必需的信息; 涉及学生身份的数据原则上采用匿名编号或替代编码处理; 学生仅可查看本人的诊断结果, 教师可查看本班数据, 研究分析仅使用脱敏后的汇总数据; 同时, 对数据访问、修改和导出行为保留日志, 并设置明确的数据保存周期与清理机制。

8.4. 研究局限与未来改进方向

本研究仍存在若干局限。其一, 当前系统更适合知识点边界清晰、错误类型较稳定的题目, 对开放性推理题、复杂证明题和跨章节综合题的适配性仍有限。其二, 本文建立了四类主要错误类型框架, 但对多重错因并发和长期迁移性误解的刻画仍较为粗略。其三, 本文目前主要属于设计与实践论证, 尚缺乏大样本、长期追踪和对照研究基础上的实证检验。未来可进一步扩充错因数据集、优化教师修正回写机制, 并在更大范围的真实教学环境中持续验证系统效果。

参考文献

- [1] 周蓓, 尹绍军. 高中数学学习中的知识误解及其矫正策略研究[J]. 教育进展, 2025, 15(8): 885-890.
- [2] 王思宇, 孙爱慧, 王喆, 郝连明. 小学与初中数学学习衔接问题的成因剖析与对策研究[J]. 教育进展, 2024, 14(1): 285-291.
- [3] 高陈欣玥, 郑富. 构建中学生错题分析系统[J]. 科学咨询, 2022(4): 1-4.