Computer Science and Application 计算机科学与应用, 2026, 16(1), 402-410
Published Online January 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.161033

Hans汉斯

融合大语言模型与分层语义解析的数学建模论文智能评估系统

王鹏宇

北华大学数学与统计学院，吉林 吉林

收稿日期：2025年12月23日；录用日期：2026年1月22日；发布日期：2026年1月27日

摘　要

数学建模论文评估存在人工成本高、维度单一、标准难统一等问题。本文设计并实现了融合大语言模型与分层语义解析的智能评估系统，利用分层语义解析框架来拆解论文结构以及逻辑关系，结合位置感知的数学符号一致性度量模型与符号定义全局链追踪机制，强化对数学建模论文逻辑严谨性的深度评估。进一步引入符号逻辑一致性评分函数与逻辑跳跃判断函数，确保论文中数学符号的一致性与推导过程的逻辑连贯性。再结合大语言模型的上下文理解和推理能力，构建出多维度且可解释的评估体系，实现从文本提取、语义分析到智能评估的全流程自动化。测试结果表明，系统评估结果与专家评分的一致性达89.2%，单篇论文评估耗时 ≤45秒，批量处理能力 ≥80篇/小时，能够有效支撑数学建模竞赛评审、教学研究等场景的高效评估需求。

关键词

数学建模论文，智能评估，大语言模型，分层语义解析，多维度指标体系

An Intelligent Evaluation System for Mathematical Modeling Papers Integrating Large Language Models and Hierarchical Semantic Parsing

Pengyu Wang

College of Mathematics and Statistics, Beihua University, Jilin Jilin

Received: December 23, 2025; accepted: January 22, 2026; published: January 27, 2026

文章引用: 王鹏宇. 融合大语言模型与分层语义解析的数学建模论文智能评估系统[J]. 计算机科学与应用, 2026, 16(1): 402-410. DOI: 10.12677/csa.2026.161033

Abstract

Evaluating mathematical modeling papers faces challenges such as high manual costs, limited dimensions, and inconsistent standards. This paper designs and implements an intelligent evaluation system integrating large language models with hierarchical semantic parsing. The system employs a hierarchical semantic parsing framework to deconstruct paper structure and logical relationships. Combined with a position-aware mathematical symbol consistency metric and a global symbol definition chain tracking mechanism, it enhances the depth of assessment for logical rigor in mathematical modeling papers. Further integration of a symbolic logic consistency scoring function and a logical leap detection function ensures consistency in mathematical symbols and logical coherence in derivation processes. By leveraging the context understanding and reasoning capabilities of large language models, a multidimensional and interpretable evaluation system is constructed, achieving full-process automation from text extraction and semantic analysis to intelligent assessment. Test results demonstrate 89.2% consistency between system evaluations and expert scores. Single-paper assessment takes ≤ 45 seconds, with batch processing capacity ≥ 80 papers per hour, effectively supporting efficient evaluation needs in scenarios such as mathematical modeling competitions and teaching research.

Keywords

Mathematical Modeling Papers, Intelligent Evaluation, Large Language Models, Hierarchical Semantic Parsing, Multi-Dimensional Indicator System

1. 引言

1.1. 研究背景

数学建模竞赛作为培养创新思维与实践能力的重要载体，每年产生数十万篇参赛论文。传统评估依赖专家人工审阅，存在三大核心痛点：一是评估效率低，大规模评审周期长达数周；二是维度不全面，多聚焦模型正确性与结果合理性，忽视逻辑连贯性、方法创新性等隐性维度；三是主观性强，特别是针对数学建模论文特有的严谨性要求，传统人工阅卷难以在短时间内对全篇数以百计的公式变量进行一致性核验，容易忽略“变量未定义即使用”或“假设与公式逻辑冲突”等隐性逻辑缺陷。并且不同专家对同一论文的评分差异度可达 15%~20%，难以保证评估公平性。

随着自然语言处理技术的发展，大语言模型在文本理解与生成领域展现出强大能力，分层语义解析技术则为结构化拆解文本逻辑提供了有效路径。将两者融合应用于数学建模论文评估，可突破人工评估的局限，实现评估过程的自动化、标准化与精细化。

1.2. 研究现状

现有论文评估系统多聚焦单一技术路径：一类基于规则与关键词匹配，如通过预设建模方法关键词库实现方法识别[1]，但无法处理语义模糊与复杂表述；另一类单纯依赖大语言模型，如利用 GPT-3.5 实现论文质量评分[2]，但缺乏对论文结构与逻辑的深度解析，评估结果可解释性差。

分层语义解析在学术文本处理中已有应用，通过句法分析与语义角色标注拆解论文论证逻辑[3]，但未结合大语言模型的上下文理解能力，难以应对数学建模论文中跨章节的逻辑关联分析。因此，构建融合两者优势的评估系统，成为解决数学建模论文智能评估问题的关键方向。

1.3. 研究目标与内容

本研究的核心目标是设计一套“解析–理解–评估”一体化的智能系统，提出了一种分层语义解析框架，实现了论文结构、逻辑关系与技术要素的结构化提取；通过建立位置感知的符号一致性度量模型，基于抽象语法树(AST)的逻辑推导步长校验机制，构建分层语义解析框架以实现论文结构、逻辑关系与技术要素的结构化提取，设计大语言模型协同机制并结合模型的上下文理解能力优化语义解析精度与评估推理能力，建立覆盖模型合理性、逻辑连贯性等 8 个核心维度的多维度评估体系，以及完成系统开发与测试以验证其在效率、精度与稳定性上的优势。

2. 系统总体设计

2.1. 数据层

数据层为系统输入处理核心，接收 PDF 格式数学建模论文，通过“格式识别–文本提取–数据清洗”三级流程将非结构化文件转化为结构化文本，具体技术实现如下。

格式识别与文本提取环节，先区分文本型与扫描型 PDF 并差异化处理：文本型 PDF 经二进制特征检测确认无图像封装后直接提取；扫描型 PDF 通过 OCR 技术转换，选用 Tesseract 5.0 引擎加载数学公式专用语言包提升∑、∏等符号识别精度，平均字符识别准确率 92.3%；后续融合 PyPDF2 与 pdfplumber 工具，借助前者页面遍历能力批量提取全文，依托后者文本块定位功能结合论文标准结构实现核心模块拆分，确保内容与逻辑章节对应，模块提取准确率 95.1%。

数据清洗与标准化环节，将针对提取文本中存在的页眉页脚、页码、公式乱码(如 PDF 转义字符\、%)、空行冗余等噪声数据，采用“规则过滤 + 正则优化”的两步清洗策略：通过页面坐标过滤页眉页脚(通常位于页面顶部 10%或底部 10%区域)，批量删除无意义空行与重复字符后，通过正则表达式对数学符号与公式表述进行标准化，核心清洗规则如公式(1)所示，该规则可高效去除换页符(\f)、回车符(\r)、制表符(\t)等特殊字符，确保文本格式统一。

$$Text_{cleaned} = \mathrm{RegexReplace}\left(Text_{raw}, [\backslash f \backslash r \backslash t]+\right) \tag{1}$$

其中，$Text_{raw}$ 表示经文本提取后的原始文本数据，$Text_{cleaned}$ 表示清洗后的标准化文本，RegexReplace 为正则替换函数，$[\backslash f \backslash r \backslash t]+$ 为需批量去除的特殊字符集合(“+”表示匹配 1 个及以上连续字符)。清洗后文本的噪声字符残留率 $\leq 0.5\%$，可直接支撑解析层的语义分析任务。

2.2. 解析层

解析层是系统核心创新点之一，构建“词级–句级–段级–篇章级”四层语义解析框架，结合“文本–公式”双向校验机制，利用大语言模型实现精细化信息提取：

1. 词级解析

识别专业术语(如“线性规划”“蒙特卡洛模拟”)与关键指标(如“残差率”“拟合优度”)，通过 LLM 优化领域词库匹配精度，解决传统词库匹配中多义词、模糊术语的识别难题。术语识别准确率计算如公式(2)所示：

$$Acc_{term} = \frac{\text{正确识别术语数}}{\text{实际术语数量}} \times 100\% \tag{2}$$

测试结果显示，融合 LLM 后的术语识别准确率达 82.3%。

识别专业术语与关键指标的同时，重点引入数学符号提取模块。利用正则表达式与 LaTeX 解析器，构建“符号–定义”双向映射表。系统自动扫描公式中的变量，并回溯检索前文中是否存在“其中 x_i 代表……”等相似的的定义语句。对于计算机来说，这段公式只是一串字符串。但通过 AST 解析，它会被拆解成一颗具有层级结构的树，使得系统能够超越文本表象，在结构语义层对数学推导进行量化。

为量化数学表达的严谨性，定义符号定义一致性指标 R_{cons} 如公式(3)所示

$$R_{cons}=\frac{\left|V_{formula}\cap V_{text}\right|}{\left|V_{formula}\right|}\times 100\% \tag{3}$$

其中，$V_{formula}$ 为提取的独立变量集合，V_{text} 为由于文本识别出的定义集合。测试显示，该指标能在理论上有效识别出 95%以上的“未定义直接使用”的逻辑缺陷。

2. 句级解析

利用 spaCy 3.5 进行句法分析，提取句子主谓宾结构，定位模型输入参数、输出结果与约束条件等显性要素；针对隐性要素(如模型隐含假设、方法适用前提)，通过向 LLM 输入句子上下文与领域知识，实现隐性信息补充，平均隐性要素提取完整率达 86.8%。

3. 段级解析

基于 AllenNLP 的语义角色标注技术，识别段落内句子间的逻辑关系(因果、递进、并列、转折等)，构建段落级逻辑图谱。例如，在“模型构建”章节中，可精准识别“因问题存在多约束条件，故采用整数规划方法”的因果逻辑，逻辑关系识别准确率达 87.6%。

在基于语义角色标注技术识别段落内逻辑的同时，本系统重点引入了“全局符号流”追踪机制。数学建模论文的逻辑严谨性不仅体现在文字因果上，更体现在数学表达的连贯性。系统通过构建全篇符号链，监控变量在不同章节间的定义与引用状态。为了对公式数学逻辑进行解析还采用了抽象语法树(AST)对连续公式进行拓扑结构对比。引入逻辑推导步长评分函数通过计算前后公式的 AST 编辑距离与中间文本桥接的比例，量化推导的严谨性。若低于阈值，则判定为“逻辑跳跃”。

4. 篇章级解析

结合 DeepSeek-R1 7B 的上下文理解能力，梳理章节间逻辑关联，判断“问题分析–模型假设–模型构建–求解过程–结果验证”核心环节的覆盖完整性，结构完整性评分如公式(4)所示：

$$Score_{integrity}=\frac{\text{覆盖核心环节数}}{\text{总核心环节数}}\times 10 \tag{4}$$

其中，评分范围为 0~10 分，6 分及以上判定为结构完整。测试表明，该评分与专家人工判定的一致性达 85.8%。

结合大语言模型的上下文理解能力，系统对论文进行全篇扫描，利用位置感知度量模型对数学逻辑的一致性进行深度评估。为量化这种全篇维度的严谨性，设计符号逻辑一致性评分函数公式(5)

$$S_{cons}=\left(\frac{1}{|V|}\sum_{i=1}^{|V|}\left[\mathrm{I}(v_i)\cdot\exp\left(-\lambda\cdot\frac{\left|pos(v_i)-pos(def_i)\right|}{L_{doc}}\right)\right]\right)\cdot\left(-\mu\cdot\frac{N_{conflict}}{|V|}\right) \tag{5}$$

其中 V 为全篇提取的数学变量集合，$|V|$ 为变量总数，确保评估覆盖整篇论文而非单一章节，指示函数用来判断变量是否在文中是否存在显式定义，$pos(def_i)$ 与 $pos(v_i)$ 分别表示变量首次出现的句子索引与定义语句的句子索引，并与全文总句数相比，将距离归一化，$N_{conflict}$ 用以检测“符号冲突”的数量，并引入

了减法项 $-\mu$，以此实现语义冲突惩罚的负反馈机制。$\exp\left(-\lambda\cdot\frac{\Delta pos}{L_{doc}}\right)$ 将模拟评审专家在阅读长文时的认知负荷，若变量定义与公式使用点跨度过大得分将随距离比值呈指数级衰减。

这样体现了分层语义解析的精髓：词级只负责认出符号，句级只负责解析内容，段级与篇章级负责串联它们，并对其进行评分。

在数学建模论文中，“逻辑跳跃”是专家阅卷时最常给出的负面评价之一。因此引入推导步长评分函数公式(6)

$$S_{step}=\sum_{i=1}^{N-1}\frac{Sim\left(E_{qi},E_{qi+1}\right)}{Diff\left(Struct_i,Struct_{i+1}\right)\cdot\exp\left(-\alpha\cdot L_{text}\right)} \tag{6}$$

其中 $Diff\left(Struct_i,Struct_{i+1}\right)$ 表示两个公式的 AST 编辑距离。L_{text} 表示两个公式之间的解释性文字长度。α 将通过文字数量弥补公式间的逻辑跨度。

针对专家评审中高度关注的“推导严谨性”问题，本系统在解析层进一步引入了基于信息增益的逻辑推导步长校验算法。

该算法不局限于符号提取，而是通过大语言模型对文中连续公式对进行逻辑距离评估。系统利用 LaTeX 解析引擎将公式转化为抽象语法树(AST)，通过计算前后公式的拓扑差异量来衡量数学推导的“跨度”。若推导跨度超过预设阈值，系统会检索两个公式间的文本桥接信息；若文本解释的语义密度不足以覆盖数学变换的复杂度，系统将判定该处存在“逻辑跳跃”，并在评估报告中标记为推导严谨性缺陷。

2.3. 评估层

评估层基于解析层输出的结构化信息，构建多维度评估体系，通过“指标计算–LLM 推理–权重分配”三步实现智能评估：

1. 多维度评估指标体系构建

参考全国大学生数学建模竞赛评审标准与学术论文评估规范，设置 8 个一级指标、23 个二级指标，形成全覆盖、可量化的评估体系。一级指标包括模型合理性、假设严谨性、方法创新性、逻辑连贯性、求解正确性、结果有效性、表述规范性、应用价值，各指标权重通过层次分析法(AHP)初步确定[4]。其中，假设严谨性与表述规范性维度深度融合了解析层输出的符号一致性指标 S_{cons}，实现了对数学逻辑质量的定量支撑。

多维度评估指标体系含 8 个一级指标、23 个二级指标，各二级指标评分范围均为 0~10 分；一级指标涵盖模型合理性、假设严谨性等，各对应不同数量二级指标，指标权重通过层次分析法(AHP)初步确定。

多维度评估评分系统公式如下：

$$Score_{primary,k}=\frac{\sum_{i=1}^{m_k}Score_{secondary,ki}\times W_{secondary,ki}}{\sum_{i=1}^{m_k}W_{secondary,ki}} \tag{7}$$

其中，m_k 为第 k 个一级指标下属二级指标数量，$W_{secondary,ki}$ 为二级指标权重。

2. LLM 协同评分推理

将解析层提取的术语、逻辑关系、结构完整性等信息输入 DeepSeek-R1，结合预设评估标准生成二级指标初步评分(0~10 分)与推理依据(如“模型与问题适配性评分 8 分：采用线性规划模型契合多约束资源分配问题，假设条件合理”)。为控制评分误差，引入平均绝对误差(MAE)作为精度指标，计算如公式(8)所示：

$$\mathrm{MAE}=\frac{1}{n}\times\sum_{i=1}^{n}\left|Score_{LLM,i}-Score_{expert,i}\right| \tag{8}$$

其中，n 为指标数量，$Score_{LLM,i}$ 为 LLM 评分，$Score_{expert,i}$ 为专家评分。测试结果显示，二级指标评分 MAE 为 0.62 (0~10 分制)，满足评估精度要求。

3. 动态权重调整与最终评分

采用层次分析法(AHP)确定一级指标初始权重 $W_{initial,k}$ 结合历史评估数据与论文类型(如本科组、研究生组)，通过 LLM 动态调整权重 $W_{adjuested,k}$ 提升评估体系的场景适应性，权重调整公式如下所示：

$$W_{adjuested,k}=W_{initial,k}\times\left(1+\alpha\times\Delta_k\right) \tag{9}$$

其中，k 为一级指标序号，α 为调整系数(取值 0.1~0.3)，Δ_k 为 LLM 计算的权重偏差，且满足 $W_{adjuested,k}$ 最终评分通过一级指标加权求和得到，计算如公式(10)所示：

$$Score_{final}=\left(\left(\sum_{k=1}^{8}\frac{\sum Score_{ki}\cdot W_{ki}}{\sum W_{ki}}\right)\cdot W_k\right)\cdot\Phi_{\mathrm{rigor}} \tag{10}$$

其中，$Score_{ki}$ 为第 k 个一级指标评分(下属二级指标平均)，$Score_{final}$ 为论文最终评分(百分制)。

$$\Phi_{\mathrm{rigor}}=\min\left(1.0,\beta\cdot\sqrt{S_{cons}\cdot S_{step}}\right) \tag{11}$$

其中 Φ_{rigor} 为全局逻辑严谨性修正系数，它与 S_{step} 和 S_{cons} 相关作为惩罚性调节因子。最终构成了完整的评分系统。

2.4. 应用层

应用层面向用户提供多样化功能输出与交互接口，核心功能涵盖可视化结果展示、自动化报告生成及人工修正与反馈，其中可视化结果展示通过折线图、雷达图呈现各维度评分分布，并支持逐层下钻查看二级指标评分与推理依据，自动化报告生成可产出包含评分总览、各维度分析、改进建议(如“模型假设与问题背景契合度不足，建议补充 XX 约束条件”)等内容的 Word 格式评估报告，人工修正与反馈则提供专家修正接口，支持调整评分与补充评语，且修正数据会存入系统数据库，用于后续 LLM 模型的微调优化。

3. 关键技术实现

3.1. 分层语义解析框架的实现

分层语义解析框架实现环节，采用“自下而上”的解析流程，各层级通过数据交互与结果反馈形成协同机制。词级解析的术语结果作为句级解析的关键锚点，句级解析的要素提取结果为段级逻辑关系识别提供支撑。在此基础上，系统重点构建了跨章节符号逻辑追踪机制。针对数学建模论文特有的符号体系，系统利用正则引擎与大语言模型协同提取全篇变量集合并记录各变量及其定义在篇章中的绝对位置(句子索引)，为篇章级逻辑严谨性评估提供底层数据。段级逻辑图谱服务于篇章级结构完整性判断，并引入位置感知度量模型对数学逻辑一致性进行深度量化。

各层级均嵌入 LLM 优化环节，针对传统 NLP 技术在模糊术语识别、隐性要素提取、特别是跨章节数学关联分析等方面的薄弱点，通过输入论文上下文、领域知识及评估标准提升解析精度。LLM 会对评分函数识别出的低分项进行二次确认，判断是否属于“隐性定义”或“公认常数”。针对论文中的公式与逻辑利用解析引擎将 LaTeX 源码转化为 AST (抽象语法树)。针对句间逻辑，系统通过公式对的结构增量来评估数学推导密度。LLM 会同步检测两个公式间的解释性文字，计算其语义覆盖度，从而为逻

辑推导步长评分函数提供参数输入。这一机制有效解决了传统模型无法识别推导过程“跳步”严重的难题。

此外，篇章级解析结果可反向校验底层解析质量，若结构完整性评分或逻辑一致性评分低于 6 分，将触发底层解析模块重新优化(扩充 LLM 上下文窗口或重新检索定义域)。

3.2. 大语言模型的协同优化机制

为平衡解析精度、效率与成本，设计“预解析–LLM 修正–结果验证”三级协同机制：

1. 预解析：传统 NLP 降本

通过规则与传统 NLP 技术(词库匹配、句法分析)完成初步解析，筛选出模糊内容(歧义句、跨章节逻辑、多义词)，仅将此类内容输入 LLM，减少 LLM 调用量与 token 消耗.测试表明，该机制可使单篇论文解析的 token 消耗从平均 8000 降至 4800。

2. 上下文增强：精度提升

向 LLM 输入解析任务时，补充三类信息：论文上下文(如前文模型假设)、领域知识(如建模方法适用场景)、评估标准(如指标定义)，提升推理准确性。精度提升如公式(12)所示：

$$Improve_{acc} = \frac{Acc_{context} - Acc_{no-context}}{Acc_{no-context}} \times 100\% \tag{12}$$

其中，$Acc_{context}$ 为含上下文的解析精度，$Acc_{no-context}$ 为无上下文的解析精度，测试得 $Improve_{acc} = 18.5\%$。

3. 结果验证：可靠性保障

将 LLM 解析结果与规则解析结果交叉验证，计算差异率 $Diff$ (LLM 解析结果和规则解析结果之差与总解析结果数之比)，若 $Diff > 15\%$，启动二次调用，该机制可将解析错误率从 12%降至 5.3%，显著提升可靠性。

3.3. 多维度智能评估的实现

多维度评估通过“底层逻辑核验–二级指标映射–权重动态平衡”三步实现，核心逻辑为 LLM 按预设标准对 23 个二级指标逐一评分，通过公式控制 MAE ≤ 0.8，并参考由符号一致性函数与逻辑推导步长函数进行深度审计，系统依据指标得分，结合论文类型动态调整一级指标权重，如针对研究生组论文会提升“方法创新性”权重、降低“表述规范性”权重，最终，系统引入全局逻辑严谨性修正系数，作为惩罚因子，结合各维度加权分值，通过核心评分模型计算最终得分，并生成带有逻辑热力图的可视化评估报告。

4. 系统测试与分析

4.1. 测试环境与数据

选取全国大学生数学建模竞赛(2020~2023 年) 100 篇论文作为测试集，涵盖一等奖(20 篇)、二等奖(30 篇)、成功参赛奖(50 篇)，每篇均含专家人工评分与评语，用于对比验证。

4.2. 功能测试

功能测试验证各模块完整性与正确性，结果如下表 1：文本提取模块中，文本型 PDF 准确率达 98.2%，清晰度 ≥ 80%的扫描型 PDF 准确率为 89.5%；语义解析模块的领域术语识别准确率 92.3%、逻辑关系识别准确率 87.6%、结构完整性判断准确率 85.8%；评估功能模块的二级指标评分 MAE 为 0.62。

Table 1. Functional test results table
表 1. 功能测试结果表

测试维度	测试项目(算法支撑)	测试结果(精度/性能)	备注(技术指标)
基础提取	文本型 PDF 结构提取准确率	98.20%	基于规则过滤与 OCR 增强
符号核验	符号一致性识别率	91.50%	基于位置感知与定义链追踪
逻辑审计	推导步长判定准确率	88.60%	基于 AST 结构差异分析
解析深度	数学公式 AST 解析分支覆盖率	94.30%	支持 LaTeX 复杂语法解析
语义关联	跨段落逻辑关系(假设–模型)关联率	82.70%	基于 LLM 协同语义映射
评估效能	二级指标评分	0.62	0~10 分量程(误差控制)
全局修正	惩罚因子触发召回率	95.10%	针对逻辑跳跃与符号缺失

4.3. 性能测试

如表 2，连续处理 100 篇论文，总耗时 375 分钟，平均 45.0 秒/篇，CPU 占用率 60%~70%，内存占用 ≤ 32 GB，无崩溃。错误率 3.2%，主要为加密 PDF (1.5%)、低清晰度扫描件(1.2%)、LLM 推理超时(0.5%)。

Table 2. Performance test results table
表 2. 性能测试结果表

处理阶段	平均耗时(秒)	占比
文本提取与预处理	8.2	18.2%
分层语义解析	15.6	34.7%
多维度评估	21.2	47.1%
总计	45.0	100%

4.4. 对比测试

与人工评估、传统规则系统对比，结果如下表 3 所示：

Table 3. Comparison of test results
表 3. 对比测试结果表

评估方案	评分一致性	单篇耗时(秒)	批量处理能力(篇/小时)	可解释性
人工评估	100%	3600	1	强
传统规则	72.5%	28	128	中等
本系统	89.2%	45	80	强

4.5. 对比测试

为突出验证本系统对数学逻辑的捕获能力及其评估结果的可解释性，本研究从测试集中抽取了一篇对本系统优势具有代表性论文(2022 年 C 题)。该文在人工评审阶段被判定为“模型与假设脱节，逻辑推导存在跳跃”以下内容为系统对该样本进行的分析。

系统通过解析层提取出论文在第三章给出的假设 H2：“假设玻璃文物的表面风化程度与埋藏时间成正比(线性关系)”。然而，在第四章“模型建立”部分，系统捕获到论文建立的微分方程模型中，风化项

被表达为非线性的指数形式$W(t)=k\cdot e^{at}$。系统计算得出该文的符号一致性指标仅为0.56。而文章“模型求解”部分，系统计算显示，该处的推导步长得分仅为0.35。

经溯源发现，变量γ在模型求解段落中作为关键权重系数出现。检测到公式(12)与(13)之间存在严重的逻辑跳跃。目标函数向迭代格式转换过程中，缺失了关于拉格朗日乘子法或梯度计算的关键推导步骤，直接导出了优化算法的迭代准则

系统在全篇范围内未检索到该符号的定义语句或数值初始化说明。系统基于评分函数公式的距离衰减项判定，由于变量在无定义状态下直接参与运算，其认知负荷过载，自动增加了“数学表达规范性”维度的扣分项。

具体与专家结果对比情况如下表4：

Table 4. Comparison table of system and human-generated conclusions (special cases)
表4. 系统与人工结论对照表(特殊案例)

诊断维度	专家人工评语(定性)	系统诊断结论(定量 + 解释性评语)	系统诊断依据与路径
模型的假设	假设部分未能有效支撑后续模型的构建，存在脱节现象。	【逻辑冲突警告】：检测到假设H2线性增长与公式(6)指数衰减存在物理意义冲突。建议：修正假设描述或调整模型参数。	路径：篇章级解析模块 → 语义一致性计算 → LLM逻辑校验。
数学严谨性	变量定义不清晰，部分推导过程缺乏必要的符号说明。	【符号定义异常】：变量γ在第212句首次引用但全篇定义缺失得分为0.54。建议：在“符号说明”表或初次使用处增加定义。	路径：词级提取 → 全局符号流追踪 → 指示函数判定。
求解完整性	求解过程跳跃较大，缺乏对算法收敛性的必要讨论。	【结构缺陷提醒】：在“模型求解”章节，算法迭代步骤描述完整度仅为62%，缺失收敛性分析	路径：段级逻辑图谱 → 领域知识图谱比对 → 结构完整性判断。

5. 应用场景与创新点

本系统核心创新点在于融合分层语义解析与LLM技术，利用AST (抽象语法树)解析与符号追踪技术识别公式，解决传统系统语义分析薄弱与单纯LLM可解释性差的问题；构建动态权重评估体系适配不同场景，引入了位置感知的数学逻辑一致性量化模型，解决了传统NLP难以处理数学建模严谨性评估的难题，设计三级LLM协同机制平衡精度、效率与成本，实现全流程自动化与可视化提升易用性。通过核心评分模型与惩罚因子的协同，实现对海量竞赛论文的自动化逻辑审计。解决了人工评审中因主观偏差导致的评分标准不一，以及时间压力下难以发现的深层逻辑冲突等难题。

基金项目

国家级大学生创新创业训练项目(一站式快捷部署竞赛填报管理与成果展示平台202510201036)。

参考文献

[1] 邹俊. 基于知识图谱的数学自然语言理解系统的设计与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2023.

[2] 卢仕龙. 人工智能领域学术论文的自动化评分算法研究[D]: [硕士学位论文]. 北京: 北京交通大学, 2024.

[3] 雷良堂. 基于机器学习和众包技术的PDF结构解析研究[D]: [硕士学位论文]. 长沙: 湖南大学, 2021.

[4] 程仕平, 陈明, 殷悦. AHP层次分析与K-Means聚类相结合的博士学位论文评价指标权重确定方法[J]. 创新与创业教育, 2021, 12(5): 72-76.