

Lpaa: 基于 LangGraph 的教案自动生成智能体

张明, 张嘉睿, 何雨晴, 张琰彬

华东师范大学数据科学与工程学院, 上海 200062

摘要

本文提出 Lpaa, 一种基于 LangGraph 构建的教案自动生成智能体 (Lesson Plan Automatic Agent)。该智能体由大语言模型 (Large Language Model, LLM) 驱动, 通过内容解析引擎、指令链生成器、前序步骤描述生成器、并行输出引擎与可控校准优化器等五个协同模块, 实现知识结构提取、教学任务分解、阶段衔接生成、模块并行处理与人机协同校准。Lpaa 以智能体 (agent) 形式运行在 LangGraph 的图结构上, 支持任务节点化、依赖约束表达及动态调度, 从而显著提升生成效率与教学内容的一致性。在自构建的多维度评估框架 pEval 下, 系统平均得分 8.48/10, 得到领域专家的高度认可。消融实验进一步验证了智能体各子模块对整体性能的关键贡献。研究表明, Lpaa 为中文教案自动生成提供了结构化、可扩展的智能体范式, 可为提升教学质量与学习体验提供有效支持。

关键词

课程教案生成; 大语言模型 (LLM); 智能体; Lpaa; pEval

中图分类号: TP39
BDR25242

文献标志码: A

doi: 10.11959/j.issn.2096-0271.预发表

Lpaa: Lesson Plan Automatic Agent Based on LangGraph

ZHANG Qiming, ZHANG Jiarui, HE Yuqing, ZHANG Yanbin

School of Data Science and Engineering, East China Normal University, Shanghai 200062, China

Abstract

This paper presents Lpaa, a LangGraph-based Lesson Plan Automatic Agent designed for intelligent generation of Chinese lesson plans. Powered by a Large Language Model (LLM), the agent coordinates five functional modules— a Content Parsing Engine, an Instruction Chain Generator, a Previous Step Description Generator, a Parallel Output Engine, and a Controllable Calibration Optimizer. These modules collaboratively support knowledge structure extraction, teaching task decomposition, stepwise contextual linkage, parallel module processing, and human-in-the-loop calibration. Operating as an agent within the graph-structured LangGraph framework, Lpaa enables task node segmentation, dependency-aware scheduling, and dynamic workflow management, thereby significantly improving generation efficiency and content consistency. Under a self-constructed multidimensional evaluation framework, pEval, the system achieved an average score of 8.48 out of 10 and received strong recognition from domain experts. Ablation studies further validate the critical contributions of each agent module to overall system performance. The results

预发表

BDR25242-1

demonstrate that Lpaa offers a structured and scalable agent-based paradigm for automatic Chinese lesson plan generation, providing effective technical support for enhancing instructional quality and learning experiences.

Key words

Lesson Plan Generation, Large Language Model (LLM), Intelligent Agents, Lpaa, pEval.

0 引言

课程教案是教学活动的重要组成部分，是教师实施有效教学的基本依据。通过合理的教案设计，教师能够明确教学目标、科学安排教学内容与时间、选择适当的教学方法与评价方式，从而提升课堂效率与教学质量^[1]。与此同时，优质教案的编写也为学生提供了清晰的学习路径，有助于其掌握知识结构、形成系统认知，并促进深度学习与能力发展^[2]。

然而，在当前教育数字化转型的背景下，传统人工撰写教案的方式正面临效率低、资源利用率不高等诸多挑战。一方面，教师在备课过程中需要花费大量时间整理教学内容、组织教学环节，工作负担沉重；另一方面，不同教师的个人经验和表达风格差异也造成了教案质量的不均衡，难以实现标准化与共享利用。因此，如何借助先进的人工智能技术，构建高效、规范、智能的教案生成系统，成为教育信息化领域的重要研究课题。

近年来，随着大语言模型（Large Language Models, LLM）的迅猛发展，其在文本生成方面展现出强大的能力，逐渐被应用于教案辅助生成任务中^[3]。这种方法显著提升了教案生成的速度与效率，具有广阔的应用前景。然而，已有研究和实践表明，当前基于LLM直接生成的教案在实际教学中存在诸多问题，例如：内容

覆盖不全、深度不够，容易遗漏关键教学重点；教学设计缺乏系统性与阶段性，难以体现由浅入深、循序渐进的教学原则；缺乏以学生为中心的教学视角，忽略了学生的认知特点与实际需求；教案语言呆板，缺乏教学感染力与实用性^[4]。因此，在实际教学场景中，单一依赖LLM直接生成的教案难以真正满足教学需求，也无法替代教师在教学设计中所体现的专业判断与教育智慧。

针对上述问题，本文提出并实现了一种面向中文教案自动生成的智能代理系统——Lpaa（Lesson Plan Automatic Agent），该系统以大语言模型为基础，结合图结构建模、模块化任务分解、人机协同优化等策略，从“内容结构抽取-流程构建-上下文衔接-并行生成-反馈优化”五个方面入手，力求在提升教案生成效率的同时，保障其教学连贯性与实际可用性。主要贡献如下：

1. 本文创新性地构建一个由五大核心模块组成的智能教案生成系统Lpaa，包括内容解析引擎、教学链生成器、前序步骤描述生成器、并行输出引擎以及可控校准优化器。通过模块间的协同运作，Lpaa有效应对当前教案生成过程中存在的各类问题，确保所生成教案具有连贯性与教育价值，为课程教案的智能生成提供了全新的解决思路。

2. 为验证系统的性能与准确性，本文设计了一个系统化评估框架——pEval，该框架从“内容质量”、“连贯性”与“教育

价值”三个维度对生成结果进行全面评估^[5]。实验结果表明，Lpaa在pEval评估框架下表现优异，课程教案平均得分高达8.48分（满分10分），并获得了多位学科专家的一致好评。这一高分评价充分体现了Lpaa在课程教案生成领域的显著优势与广阔应用前景。

1 相关工作

随着人工智能技术的不断发展，尤其是大语言模型（Large Language Model, LLM）在自然语言处理领域的广泛应用，基于智能代理的教学辅助工具逐渐成为教育信息化的重要研究方向^[6]。其中，教案作为教学准备的核心环节，已成为诸多研究者尝试引入智能化生成机制的重点目标。

1.1 教案生成的传统方法与自动化探索

长期以来，教案的设计主要依赖教师个人的经验积累与教学实践。这种“手工撰写+经验导向”的方式虽然具有一定灵活性与个性化优势，但存在效率低、标准不统一、质量参差不齐等问题。在教育资源共享和教学质量监控日益重要的今天，这一问题尤为突出。

为解决这一困境，近年来出现了一些利用人工智能手段辅助教案编写的研究与产品探索。早期系统如MagicSchool^[7]平台，通过简单关键词提示生成基础教案大纲；后续如SchoolAI^[8]、Curipod^[9]等工具则引入检索增强机制和多模态设计理念，尝试结合教学目标、课标知识点与教具资源，提升生成结果的教学相关性与完整性。

尽管这类工具在生成效率上取得了一定进展，但生成内容普遍存在以下问题：

内容覆盖不全：重要知识点缺失，教学目标泛化；结构逻辑薄弱：教学环节之间缺乏递进与呼应；缺乏教育理论支撑：很难体现启发式教学、以学定教等现代教学理念；泛化模板化严重：尤其在学科交叉或复杂情境中，难以适配多样教学需求^[10]。这些问题表明，当前主流方法多停留在“文本填空式”的自动生成阶段，缺乏对教学任务流程的深入建模与多层次内容理解，限制了自动教案系统的实用价值。

1.2 LLM作为教学智能体的角色

为了更好地发挥LLM的潜力，一些研究者开始尝试将其视作“智能体”（Intelligent Agent）使用，即不再将LLM仅作为生成工具，而是作为能感知上下文、推理任务逻辑并执行多阶段教学规划的“教学搭档”^[11]。

这一思路下，LLM可模拟“教师”角色，具备以下能力：结合课程内容、教学目标与学情分析，实现定制化教学活动设计；引入记忆机制^[12]与教学进度追踪，实现动态调整的教学任务链；针对不同学习者特征，生成更具个性化的教学内容与互动设计。在此背景下，具备图结构建模能力的LangGraph框架被引入教案生成流程中^[13]。其通过图神经结构组织任务节点，支持多个教学环节并行生成与跨步骤信息流转，大幅提升了教学设计的结构完整性与执行效率。例如，针对多个知识点并开展讲解、活动设计与评价安排，LangGraph都能提供高效有序的生成路径。

1.3 LLM在教案质量评价中的作用

除生成端探索外，LLM也被用于教案质量的自动评估任务^[14]。近年来，研究者

提出了包括 Auto-Arena^[15]、G-Eval^[16]等在内的一系列智能评价框架，通过多模型协同讨论、评审投票等方式，对生成教案的完整性、逻辑性、创新性等维度进行评估，从而降低人工评审成本并提升客观性。

具体到教案场景中，LLM 作为评价工具能够：识别教学目标与课时内容是否匹配；分析教学逻辑是否清晰、过渡是否自然；评价语言表达是否规范、是否具启发性；针对不同学段和学科，判断教学策略是否合理。这些研究表明，LLM 不仅在教案生成方面具有广阔应用空间，在教案评价与优化环节同样展现出强大潜力。

2 Lpaa

为解决现有教案自动生成系统中存在的效率低、结构不清、内容缺乏教学逻辑的问题，本文提出 Lpaa —— 一个基于大语言模型和图结构驱动模块化中文教案生成系统。Lpaa 从教案生成任务的本质出发，将其拆解为“内容结构识别—教学链路构建—步骤衔接说明—并行内容生成—反馈驱动优化”五个阶段，通过五大核心模块协同工作，最终实现高效、可控、富有教育价值的教案自动构建。

2.1 内容解析引擎

教材或教学文本往往存在非结构化、层级不明确、格式不统一等问题，若直接输入大型语言模型，会导致关键知识点遗漏、结构混乱或生成内容偏离教学要求。内容解析引擎通过对原始文本进行结构化建模，将其转换为包含章节层级、概念结构与核心知识点的语义表示，为后续任务分解与指令生成提供高质量输入。

设教材全文为 M ，章节定位信息为 L ，内容解析引擎的目标是从指定章节文本中抽取结构化语义表示 S ，包括层次结构、标题内容、关键术语等信息，其过程可形式化表示为：

$$S = \text{Analyze}(\text{LocateSection}(M, L)) \quad (1)$$

为实现上述目标，内容解析引擎采用基于层级结构分析^[18]与模式匹配^[19]的组合算法。首先，对教材内容进行预处理，从中提取出可能构成目录的段落信息；随后，通过分析其格式特征与结构模式，识别不同层级的目录节点。该算法的具体步骤如下：

2.2 指令链生成器

算法1:教材内容解析算法

Input: Textbook content M

LocationInfo L

Output: HierarchicalStructure S

- ① SectionText $T0 \leftarrow \text{LocateSection}(M, L)$
- ② PreprocessedText $T \leftarrow \text{Preprocess}(T0)$
- ③ LineList $L_list \leftarrow \text{SplitIntoLines}(T)$
- ④ TreeRoot $S \leftarrow \text{CreateRootNode}(\text{"Chapter"})$
- ⑤ Stack $\text{Stack} \leftarrow []$
- ⑥ for $i = 1$ to $L_list.length$ do
- ⑦ Line $line \leftarrow L_list[i]$
- ⑧ Level $level \leftarrow \text{IdentifyLevelByPattern}(line)$
- ⑨ Content $content \leftarrow \text{ExtractContent}(line)$
- ⑩ Node $node \leftarrow \text{CreateNode}(content, level)$
- ⑪ while $\text{Stack.length} > level$ do
- ⑫ Stack.pop()
- ⑬ end while
- ⑭ Stack[$level-1$].AddChild($node$)
- ⑮ Stack[$level$] $\leftarrow node$
- ⑯ end for
- ⑰ return S

教学设计是一项具有强依赖关系的多步骤任务，不同教学活动之间存在知识递进、逻辑顺承及先后依赖。若直接让LLM一次性生成完整教案，往往会出现步骤错乱、遗漏必要活动等问题。指令链生成器通过构建任务依赖图并自动完成教学目标的拆解，保证系统能够以智能体方式分步骤执行任务，从而获得更稳定、可解释且结构化的教案输出。

设教学任务 T 表示为一个需要被分解的复杂目标，其可以表示为多个子任务的集合：

$$T = \{T_1, T_2, \dots, T_n\} \quad (2)$$

其中， T_i 表示具体的第 i 个子任务。指令链 I 是子任务的有序集合，定义如下：

$$I = \{I_1, I_2, \dots, I_n\} \quad (3)$$

其中， I_i 是对子任务 T_i 的详细描述，包含操作指导与预期达成目标。其生成过程可表示为：

$$I_i = \text{Generate}(T_i, S) \quad (4)$$

为生成具备逻辑性的教学指令链，指令链生成器采用了一种基于任务优先级与依赖关系的排序算法^[20]。该算法的具体流程如下：

此外，为确保生成的教学指令链具备良好的可读性与逻辑连贯性，系统会对每一个步骤进行评估与优化。如果某一步骤的描述不清晰或存在逻辑问题，系统将自动对其进行调整与修正，从而保障最终生成的指令链具备较高的质量。

2.3 前序步骤描述生成器

在智能体执行多步任务的过程中，每一步都依赖前一步的生成结果。若缺乏对前序步骤的明确描述，LLM 在多轮任务执

算法2:教学指令链生成算法

Input: Teaching task T

Output: Instruction chain I

```

① Sub-task Set Tasks  $\leftarrow$  TaskDecomposition( $T$ )
② Dependency Graph  $G \leftarrow$  CreateDirectedGraph()
③ for each sub-task  $T_i$  in Tasks do
④   for each sub-task  $T_j$  in Tasks do
⑤     if ExistsDependency( $T_i, T_j$ ) then
⑥        $G.AddEdge(T_j, T_i)$ 
⑦     end if
⑧   end for
⑨   Priority[ $T_i$ ]  $\leftarrow$  EvaluatePriority( $T_i$ )
⑩ end for
⑪ SortedTasks  $\leftarrow$  TopologicalSort( $G$ )
⑫ for each Same-Level Task Group in SortedTasks do
⑬   Group  $\leftarrow$  SortByPriority(Group)
⑭ end for
⑮ Instruction Chain  $I \leftarrow []$ 
⑯ for each sub-task  $T_i$  in SortedTasks do
⑰   Instruction  $\leftarrow$  GenerateInstruction( $T_i$ )
⑱   Quality Score Score  $\leftarrow$  EvaluateQuality(Instruction)
⑲   if Score < Threshold then
⑳     Instruction  $\leftarrow$  OptimizeInstruction(Instruction)
    end if
    I.Add(Instruction)
  end for
return I

```

行中容易产生上下文丢失或内容漂移，导致整体流程断裂。前序步骤描述生成器通过生成简洁、结构化的前序摘要，使任务节点之间形成稳定的语义连接，是保证LangGraph流程一致性的重要组件。

设 A_i 表示第 i 个教学步骤的操作， D_{i-1} 表示前一教学步骤的描述内容。前序步骤描述生成器的目标是生成当前步骤所需的上一步骤描述 D_i 。该模块采用基于深度学习的文本生成算法^[21]，结合注意力机

制^[22]与 Seq2Seq 模型^[23]，以确保所生成的描述性内容具备准确性与较高的语义连贯性。该过程可表示为：

$$D_i = \text{Generate_Discription}(A_i, D_{i-1}) \quad (5)$$

如图 1 所示，前序步骤描述生成器的工作流程包括左侧的上下文处理模块和右

侧的任务分解与并行处理模块。系统首先对历史内容进行分析，再通过描述生成器生成描述内容。该模块将任务分解为多个关键点，并由并行输出引擎进行整合，最终输出完整的描述内容。此类并行处理机制显著提升了描述生成的效率与质量。

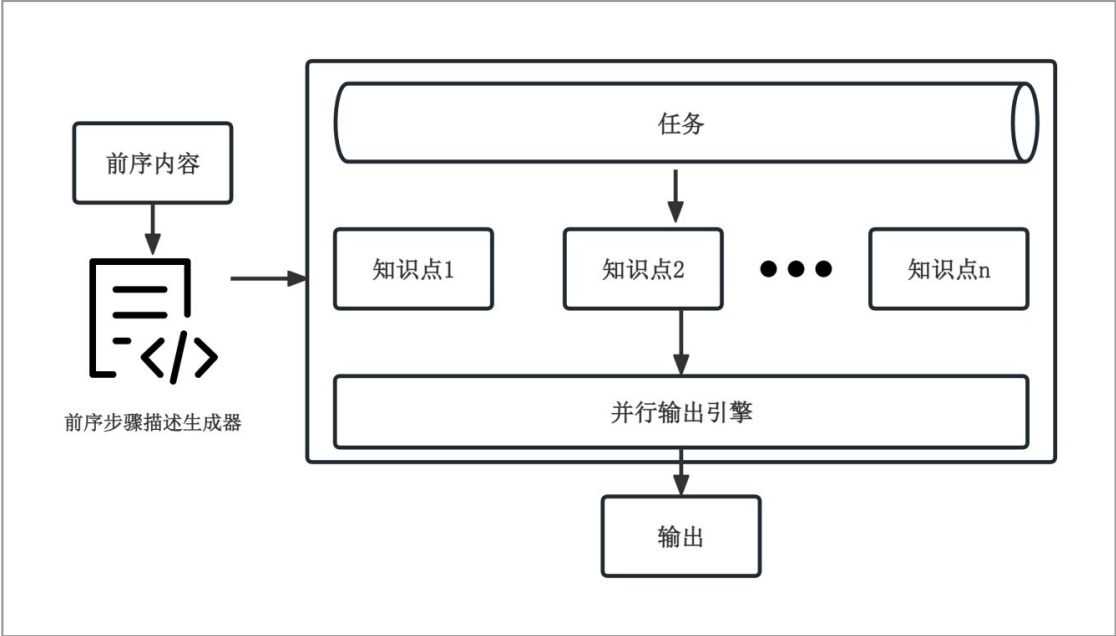


图1 前序步骤描述生成器与并行输出引擎架构图

2.4 并行输出引擎

部分教学活动或知识块之间具有弱依赖关系，可被并行生成。若采用传统串行方式生成，系统在复杂课程内容下效率较低，且容易因长链推理导致模型内容漂移。并行输出引擎利用 LangGraph 的并行调度能力，将独立任务分配到不同节点并行执行，在保持内容一致性的前提下显著提升生成效率。

并行输出引擎利用 LangGraph 的能

力，对每一内容片段进行并行化的流式输出处理^[24]。该模块通过并行处理显著提升了输出效率，能够高效生成与呈现课程提纲内容，从而增强用户的使用体验。

设 O 表示完整的课程提纲内容。并行输出引擎首先将其划分为若干个相互独立的子模块 $o = \{o1, o2, ..., on\}$ ，每个子模块对应一个知识点或教学任务。该分块过程可表示为：

$$\{o1, o2, ..., on\} = \text{Chunk}(O) \quad (6)$$

随后，并行输出引擎采用动态任务调度与资源分配算法，结合多线程或分布式计算方式，对每一个子模块 oi 进行独立生成与优化^[25]。输出过程可表示为：

$$pi = \text{Output}(Qi) \quad (7)$$

其中， pi 为子模块 Qi 的最终输出结果。该并行化流式输出机制在保证生成质量的同时，显著提升了整体生成效率，有效增强了用户的使用体验。

2.5 可控校准优化器

自动生成的教案往往需要根据教学场景、教师偏好或学段特征进行调整。若缺少人机协同的微调机制，系统生成的内容难以满足实际教学需求。可控校准优化器通过引入可控参数、人机交互式反馈以及局部重写机制，使系统具备可解释性与可修正性，增强了教案的可用性与落地性。在课程提纲初步生成完成后，可控校准优化器提供人机交互功能，使用户能够基于反馈意见实时调整课程提纲内容。该模块可快速响应用户的修改建议，优化课程提纲的准确性与实用性，确保生成的课程提纲更契合教学需求与教育原则。

系统提供界面或 API，供用户（如教师或教育专家）对课程提纲内容进行评估与反馈。用户可以指出课程提纲中的错误、遗漏或改进空间，并提出具体的修改建议。可控校准优化器接收用户反馈，并基于该信息实时调整课程提纲内容。该调整过程可表示为：

$$O' = \text{Calibrate}(O, F) \quad (8)$$

其中， O 表示原始生成的课程提纲内容， F 表示用户的反馈信息， O' 为校准后的输出结果。该过程支持多轮迭代更新，直至内容达到用户满意度或满足教学目标。

本模块的设计理念参考了已有研究中关于“人机协同校准”和“可控生成”的通用框架^[26]，但并非直接复用某一具体方法，而是结合教案生成的任务特征进行了工程化适配与扩展，从而形成适配教学场景的完整智能体式优化机制。

通过可控校准优化器，Lpaa 不仅在初次生成阶段能够保证课程提纲的整体质量，还能够根据实际教学需求持续迭代生成结果，使最终输出更契合教学原则与使用场景，显著提升系统的实用性与可落地性。

2.6 Lpaa 整体架构及协同运作

Lpaa 的整体架构如图 2 所示。五个核心模块通过数据流与控制流紧密协作，确保课程教案生成过程既高效又具备质量保障。其中，内容解析引擎为后续模块提供结构化的课程知识内容；教学链生成器与前序步骤描述生成器共同保障教学任务执行的逻辑连贯性与前后衔接合理性；并行输出引擎通过并行处理机制显著提升生成效率；而可控校准优化器则负责对生成内容进行验证与优化，确保其连贯性与教育有效性。

通过上述模块化设计，Lpaa 不仅有效提升了课程教案生成的效率，还在内容生成过程中保持了较高的科学严谨性与教学指导价值。

3 实验验证

3.1 数据集

我们构建了 PK12 数据集，该数据集主要包含语文与数学领域的 K12 阶段教材的文本文件。所有教材文本均经过注释、清洗与规范化处理，并由内容解析引擎生

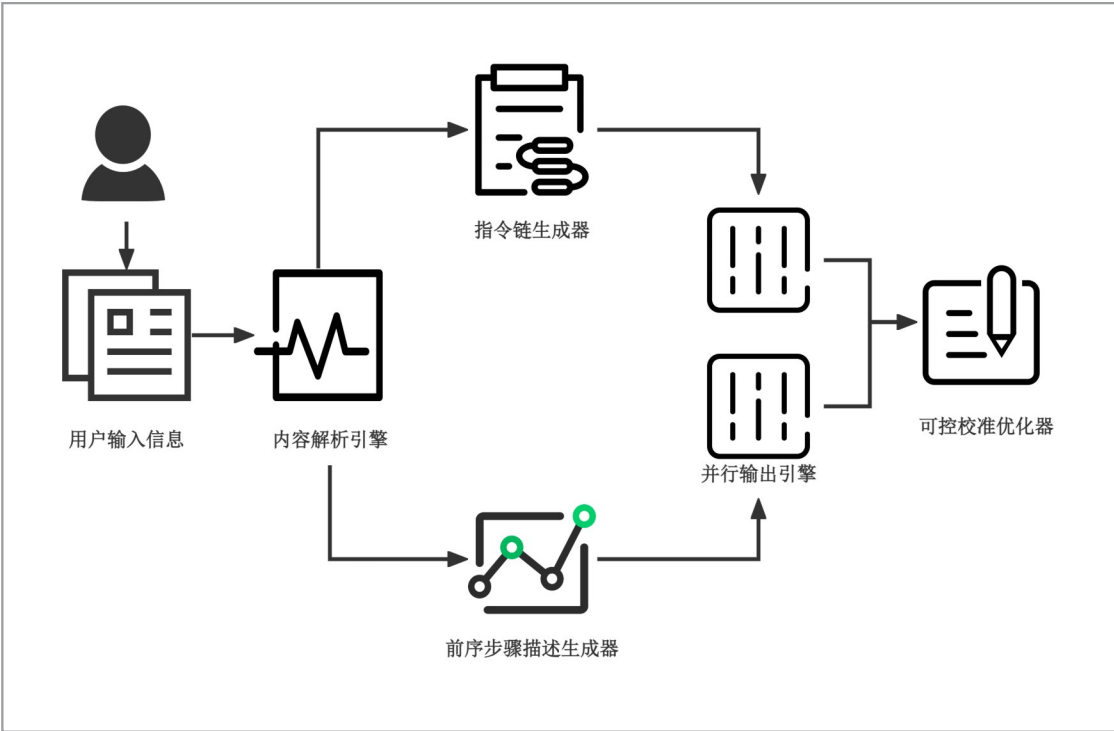


图2 Lpaa框架的整体架构与工作流程

成了对应的结构化目录信息^[27]。数据集的统计信息汇总如表1所示。

学科	表1 PK12 数据集的统计信息			
	教材		目录	
	教材页数	字符数	目录数量	字符数
语文	140.4	112886.4	5.6	733.2
数学	194.8	136938.6	3.6	909.4

3.2 实验细节

我们选取了以下基线方法进行对比：

常规方法 (Common Method)：直接使用 MagicSchool 平台生成课程教案。该方法代表了当前业界常用的自动化教案生成方式。

专业方法 (Professional Method)：利用 GPT-o3^[28]生成课程教案，随后由华

东师范大学的学科专家进行人工修订。该方法融合了先进的人工智能技术与专业教育者的经验，代表了当前领域的最新研究水平。

Lpaa 的实现基于两个模型：Qwen2.5-14B-Instruct 与 Qwen2.5-72B-Instruct^[29]。每次生成过程中，系统的输入包括学科、年级、课时数以及章节信息，随后由 Lpaa 框架自动生成教案。该设计旨在评估不同模型在生成高质量教案方面的表现。

3.3 评估指标

PPL (困惑度)：使用 Qwen2.5-32B 模型计算课程教案的平均困惑度。较低的困惑度得分表示文本流畅度更高^[30]。

Rouge-L：一种基于最长公共子序列 (LCS) 的评估指标，用于衡量生成文本与

参考文本之间的相似度。在本研究中，我们将“专业方法”作为参考文本^[31]。

Coverage（覆盖度）：评估生成的课程教案中涵盖了多少关键主题和教学目标。该指标通过检查教案是否包括预定义的关键知识点和教学目标，来衡量内容的全面性。

CLI（认知负荷指数）：认知负荷指数是一种衡量教学材料对学生认知负担的指标。它通过分析课程教案的语言复杂性、信息密度和结构清晰度，评估学生在学习

过程中可能面临的认知压力。较低的CLI值表明教案更易于理解和学习，有助于减轻学生的学习负担^[32]。

pEval：我们引入的综合评估框架，使用GPT-o3作为评估模型，从内容、连贯性和教育价值三个维度评估教案的质量，如表2所示。该框架填补了教案生成领域缺乏无参考评估的空白。pEval 框架提供0到10的数值评分，并为每个维度的评估结果提供详细的评判理由。pEval 的详细信息和实现可以参见附录。

表2 pEval 评估框架中的评估维度与评分标准

维度	评分标准
内容	内容全面深入,文字表述清晰明了。
连贯性	结构逻辑清晰,内容循序渐进地展开。
教育价值	以学生为中心,教学内容由浅入深,符合学生的认知规律,融合理论知识与现实生活应用。

3.3 实验结果

针对语文和数学科目进行了实验，结果如下：

基于上述数据集、方法和评估指标，

表3 Lpaa与基线模型在语文中的比较

配置	PPL↓ (avg.)	Rouge-L↑ (avg.)	Coverage↑ (avg./%)	CLI↓ (avg.)	pEval↑ (avg.)
Common	14.75	0.13	28.4	8.52	6.15
Professional	8.31	\	88.9	2.08	9.28
Lpaa-72b	12.87	0.762	77.4	3.12	8.69
Lpaa-14b	15.18	0.724	74.2	3.48	8.38

实验结果表明，Lpaa 系统在中文教案自动生成任务中展现出出色的性能，尤其是在采用 Qwen2.5-72B-Instruct 模型驱动下，生成质量已接近专业教师修订后的

教案水准。系统在语文学科的困惑度为12.87，数学为13.67，明显优于通用方法，说明其生成语言更为自然流畅；在内容相似度方面，ROUGE-L 分别达到

表4 Lpaa与基线模型在数学中的比较

配置	PPL↓ (avg.)	Rouge-L↑ (avg.)	Coverage↑ (avg./%)	CLI↓ (avg.)	pEval↑ (avg.)
Common	15.63	0.12	12.7	8.41	6.19
Professional	9.12	\	87.6	2.42	9.21
Lpaa-72b	13.67	0.741	70.3	3.27	8.55
Lpaa-14b	14.42	0.671	67.9	3.15	8.27

0.762（语文）和 0.741（数学），与人工修订教案保持较高一致性。

在教学目标覆盖方面，Lpaa 教案涵盖了大部分关键知识点与教学活动，语文覆盖率为 77.4%，数学为 70.3%，远超模板生成方法；此外，认知负荷指数（CLI）数值较低，语文为 3.12，数学为 3.27，表明生成教案更符合学生理解规律，有助于减轻学习负担。综合评价体系 L-Eval 的得分也进一步验证了系统的整体表现，语文平均得分为 8.69，数学为 8.55，说明其在内容完整性、教学逻辑性及教育价值方面均具备较高水平。

值得注意的是，采用 Qwen2.5-14B-Instruct 模型的 Lpaa 系统在保持较小模型规模的同时，依然取得了接近的表现。虽然困惑度略有上升（语文为 15.18，数学为 14.42），但整体表现仍显著优于常规方法。在内容相似度、覆盖率及 CLI 等指标上亦表现稳定，进一步验证了其生成教案的实用性与稳定性。

尽管两个模型的表现差异微乎其微，Lpaa 在使用 Qwen2.5-14B-Instruct 模型时提供了显著的部署优势，可能降低硬件和运营成本。此外，Lpaa 框架支持 Qwen2.5-14B-Instruct 模型的本地部署，有效保护用户数据隐私^[33]，减少数据泄露的风险。Lpaa 能够在降低成本的同时有效缩小性能差距，提供高质量的教案生成，展现了其在自动化教案生成方面的高效、

经济和安全的实际应用价值。

4 实验分析

4.1 消融实验

为了深入评估 Lpaa 框架中各个模块对课程教案生成质量的影响，我们使用 Qwen2.5-14B-Instruct 模型进行了消融实验。以语文学科为例，我们分别移除内容解析引擎（模块 1）、前序步骤描述生成器（模块 3）和可控校准优化器（模块 5）模块，测试系统性能。实验结果如表 5 所示：

表5 Lpaa消融实验结果

配置	PPL (avg.)	Rouge-L (avg.)	Coverage (avg./%)	pEval (avg.)
w/o 模块1	15.91	0.51	49.12	7.7
w/o 模块3	15.78	0.54	44.88	7.6
w/o 模块5	15.84	0.61	72.23	8.1
Lpaa-14b	15.18	0.724	74.2	8.39

消融实验结果表明，Lpaa 框架中的每个模块对生成的教案质量均有显著贡献。具体效果如下：

模块 1（内容解析相关模块）：移除模块 1 后，覆盖度降低了 25.08%（从 74.2% 降至 49.12%），突显了其在提取关键主题

和构建知识框架方面的重要性。Rouge-L 值下降了 0.214（从 0.724 降至 0.51），表明文本流畅性和内容契合度显著降低。pEval 评分下降了 0.69 分（从 8.39 降至 7.7），进一步确认了该模块对教案整体质量的重要作用。

模块 3（前序步骤描述生成相关模块）：移除模块 3 后，覆盖度下降了 29.58%（从 74.2% 降至 44.88%），显示其在确保教案完整性和连贯性方面的重要贡献。Rouge-L 值下降了 0.184，表明文本逻辑性和细节丰富度有所下降。pEval 评分下降了 0.79 分，进一步说明了该模块在提升整体生成质量中的关键作用。

模块 5（可控校准优化相关模块）：移除模块 5 后，覆盖度仅下降了 1.97%（从 74.2% 降至 72.23%），表明该模块在细粒度准确性和生成稳定性方面发挥了补充作用。Rouge-L 值下降了 0.114，pEval 评分下降了 0.29 分，说明该模块虽影响较小，但依然对最终质量有正面提升效果。

Lpaa 框架中的每个模块都对生成的教案质量起到了显著作用，各个模块在不同方面发挥着重要作用。内容解析引擎在提取关键主题和构建知识框架方面表现突出；前序步骤描述生成器对确保教案的完整性和连贯性至关重要；可控校准优化器在确保教案的准确性和可靠性方面起着重要作用。

4.2 一致性评估

pEval 与人工偏好的一致性。尽管 Chen 等人^[34]强调了 LLM 在各种生成任务中展现的接近人类的辨识能力，但评估 LLM 评价结果与人工评价结果的一致性仍然至关重要。这一观点得到了 Laskar 等人^[35]的研究支持，他们发现 LLM 在复杂任

务中的评估能力可能不足。图 3 展示了 Qwen2.5-14B-Instruct 和语文科目作为示例的专家评分与 LLM 评分之间的相关性，表明 pEval 与人工偏好的一致性。该图表明，pEval 不仅能有效地模拟专家的评估标准，还在多个维度上保持与人类偏好的一致性，进一步验证了其作为自动化教案评估工具的可靠性和实用性。

表 6 专业人工评分与大语言模型评分在不同维度下的对比结果

评价维度	专家平均 评分	pEval 评分	相关性 (ρ)
内容	8.6	8.4	0.89
连贯性	8.2	8.3	0.91
教育价值	8.7	8.6	0.87

5 总结

本文提出了一种面向中文教案自动生成任务的智能系统——Lpaa (Lesson Plan Automatic Agent)。该系统融合大语言模型生成能力与模块化结构设计，围绕内容解析、指令链生成、操作描述优化、并行输出和可控校准优化五个核心步骤，构建出高效、可控、富有教育价值的教案自动生成流程。

为验证系统效果，本文构建了教案数据集 PK12，涵盖语文与数学等典型学科，并设计了多维度评估指标体系 pEval，从语言流畅性、教学覆盖度、认知负荷、教学逻辑及教育价值等方面全面评估系统性能。实验结果表明，Lpaa 在内容质量、教学逻辑性与教育适配度方面均显著优于传统模板化或直接生成方法，接近专业教师

撰写教案水平。同时，通过模块消融实验与人工评分一致性分析，进一步验证了系统设计的合理性与评估方法的有效性。

尽管本研究取得了积极成果，但仍存在一些局限性，值得在后续工作中进一步优化：多样化学科适配性有待拓展：当前系统主要在语文与数学场景中实验验证，尚未广泛适配跨学科项目式教学等复杂类型。多模态生成能力尚不完备：目前系统侧重于文本教案生成，尚未集成如教学板书图示、流程图、实验视频等多模态教辅要素，限制了系统的课堂实用性与互动性。个性化与区域性教学场景适应仍有挑战：不同地区、学校及教师对教案风格与课程资源的偏好不同，如何引入个性化画像机制，实现真正“因人施教”的教案定制，是下一步研究重点。

未来，我们将从以下方向展开工作：拓展教案生成模型的学科覆盖范围，引入更多年级段与课程类型；探索图文多模态教案生成机制，增强可视化输出能力；引入教师知识图谱与教学风格建模，实现个性化教案推荐；推动系统在教学一线的真实应用测试，采集反馈数据，持续优化生成质量。

综上所述，Lpaa 在中文教案自动生成任务中表现出良好的应用潜力。通过模块化系统设计与多维度评估实验，我们验证了各模块在提升内容质量、教学逻辑性与适用性方面的作用。然而，该系统仍有待在更多学科场景、多模态内容以及个性化教学需求下进一步验证和完善。未来的研究将重点关注扩展学科覆盖、增强多模态输出能力、引入个性化建模机制，并在真实教学场景中持续收集反馈，以推进系统的实际应用价值与稳定性评估。

参考文献：

[1] Wang Y. Research on the Evolution of Core Literacy in Chinese Language[D]. Sichuan Normal University, 2023. DOI: 10.27347/d.cnki.gssdu.2023.000274.

[2] Liang J. The Century-Long Evolution of Core Literacy in Secondary School Mathematics in China and Its Implications[D]. Guangzhou: Guangzhou University, 2024.

[3] Wang T, Zhan Y, Lian J, et al. LLM-powered Multi-agent Framework for Goal-oriented Learning in Intelligent Tutoring System[EB/OL]. arXiv: 2501.15749, 2025.

[4] Liang X, Wang H, Wang Y, et al. Controllable text generation for large language models: A survey[EB/OL]. arXiv: 2408.12599, 2024.

[5] Hu T, Zhou X H. Unveiling LLM evaluation focused on metrics: Challenges and solutions[EB/OL]. arXiv: 2404.09135, 2024.

[6] Chu Z, Wang S, Xie J, et al. LLM Agents for Education: Advances and Applications[EB/OL]. arXiv:2503.11733, 2025.

[7] Foster A, Khazanchi P, Khazanchi R. MagicSchool. ai: A Personal Assistant [C]//Society for Information Technology & Teacher Education International Conference. AACE, 2024: 74 - 80.

[8] SchoolAI[EB/OL]. <https://schoolai.com/>.

[9] Curipod[EB/OL]. <https://curipod.com/>.

[10] Lin M P C, Chang D, Hall S, et al. Preliminary Systematic Review of Open-Source Large Language Models in Education[C]//International Conference on Intelligent Tutoring Systems. Springer, Cham, 2024: 68 - 77.

[11] 姜春宇, 韩晓璐, 王超伦. 智能体技术演进: 场景分析、关键技术与发展趋势[J]. 大数据, 2025. JIANG C Y,HAN X L,WANG C C L.

- Evolution of agent technology: scenario analysis, key technologies and development trends[J]. BIG DATA RESEARCH, 2025.
- [12] Zeng R, Fang J, Liu S, et al. On the Structural Memory of LLM Agents[EB/OL]. arXiv:2412.15266, 2024.
- [13] Wang J, Duan Z. Agent AI with Lang-Graph: A Modular Framework for Enhancing Machine Translation Using Large Language Models[EB/OL]. arXiv: 2412.03801, 2024.
- [14] Peng J L, Cheng S, Diao E, et al. A Survey of Useful LLM Evaluation[EB/OL]. arXiv:2406.00936, 2024.
- [15] Zhao R, Zhang W, Chia Y K, et al. Auto-Arena: Automating LLM Evaluations with Agent Peer Battles and Committee Discussions[EB/OL]. arXiv: 2405.20267, 2024.
- [16] Liu Y, Iter D, Xu Y, et al. G-Eval: NLG Evaluation Using GPT-4 with Better Human Alignment[EB/OL]. arXiv: 2303.16634, 2023.
- [17] Kwan W C, Zeng X, Jiang Y, et al. MT-Eval: A Multi-turn Capabilities Evaluation Benchmark for Large Language Models[EB/OL]. arXiv:2401.16745, 2024.
- [18] Shahrabi Moghadam F. Introduction to the Analytical Hierarchy Process (AHP) [J]. Zist Sepehr Student Magazine, 2024.
- [19] Nicaud C, Pivoteau C, Vialette S. Branch Prediction Analysis of Morris-Pratt and Knuth-Morris-Pratt Algorithms[EB/OL]. arXiv:2503.13694, 2025.
- [20] Pearce D J, Kelly P H J. A dynamic topological sort algorithm for directed acyclic graphs[J]. Journal of Experimental Algorithmics (JEA), 2007, 11: 1.7-es.
- [21] Yu W, Zhu C, Li Z, et al. A Survey of Knowledge-Enhanced Text Generation [J]. ACM Computing Surveys, 2022, 54 (11s): 1 - 38.
- [22] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[C]//Advances in Neural Information Processing Systems. 2017, 30.
- [23] Li Z, Cai J, He S, et al. Seq2seq dependency parsing[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 3203 - 3214.
- [24] LangChain AI. Streaming[EB/OL]. <https://langchain-ai.github.io/langgraph/concepts/streaming>, 2024-05-11.
- [25] webup: notebooks[EB/OL]. <https://github.com/webup/notebooks/tree/main>, 2024-05-11.
- [26] Bradshaw J M, Feltovich P J, Johnson M. Human - agent interaction[M]//The handbook of human-machine interaction. CRC Press, 2017: 283 - 300.
- [27] SmartEdu. tchMaterial[EB/OL]. <https://basic.smartedu.cn/tchMaterial>, 2024-05-11.
- [28] El-Kishky A, Wei A, Saraiva A, et al. Competitive Programming with Large Reasoning Models[EB/OL]. arXiv: 2502.06807, 2025.
- [29] Yang A, Yang B, Zhang B, et al. Qwen2.5 Technical Report[EB/OL]. arXiv: 2412.15115, 2024.
- [30] Hu Y, Huang Q, Tao M, et al. Can Complexity Reflect Large Language Model's Ability in Long Text Understanding?[EB/OL]. arXiv:2405.06105, 2024.
- [31] Lin C Y. ROUGE: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74 - 81.
- [32] Sweller J. Cognitive Load Theory[M]//Psychology of Learning and Motivation. Vol. 55. Academic Press, 2011: 37 - 76.
- [33] Guizani S, Mazhar T, Shahzad T, et al. A systematic literature review to implement large language model in higher education: issues and solutions[J]. Dis-

cover Education, 2025, 4(1): 1 - 25.

[34] Chen D, Chen R, Zhang S, et al. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark[C]//Forty-first International Conference on Machine Learning. 2024.

[35] Laskar M T R, Alqahtani S, Bari M S, et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 13785 - 13816.

作者简介



张昱明（1999-），男，华东师范大学数据科学与工程学院硕士生，主要研究方向为大语言模型应用、计算教育学。



张嘉睿（2000-），男，华东师范大学数据科学与工程学院硕士生，主要研究方向为大语言模型、计算教育学。



何雨晴（2001-），女，华东师范大学数据科学与工程学院硕士生，主要研究方向为大语言模型、计算教育学。



张琰彬（1983-），女，华东师范大学数据科学与工程学院工程师，主要研究方向为计算教育学、软件工程。

收稿日期: XXXX-XX-XX

通信作者: 张琰彬, ybzhang@dase.ecnu.edu.cn

基金项目:

Foundation Items: