

一种基于大语言模型的学生能力达成度智能评估方法

刘阳俊聪, 丁维龙, 张心玮, 伦亚森, 高瑞泽

(北方工业大学 人工智能与计算机学院, 北京 100144)

摘要: 当前, 能力达成度评估方法仍存在两方面问题: ① 关键能力要素依赖教师主观构建, 存在覆盖不全面、标准不一致的情况; ② 能力达成度评价视角单一, 缺乏对教学环节多过程的细粒度分析。为此, 提出一种融合大语言模型的学生能力达成度智能评估方法。首先, 针对能力要素提取设计提示词模板, 借助大语言模型自动从课程大纲中提取关键能力要素, 实现操作便捷性与语义一致性; 其次, 针对评估过程构建“题目—能力要素”映射与能力权重模型, 确定学生答题多维度量化能力要素的达成程度。将模型计算结果与教师人工评估结果进行相似度比较发现, 所提方法在准确性与实用性方面表现良好, 为数据驱动的教学评估提供了可行路径。

关键词: 大语言模型; 学生能力达成度; 教学评估; 能力映射与量化分析; 能力建模

DOI: 10.11907/rj.k.251522

扫描二维码阅读全文: 

中图分类号: TP18

文献标识码: A

文章编号: 1672-7800(XXXX)0XX-0001-09

An Intelligent Assessment Method for Students' Ability Achievement Based on Large Language Models

LIU Yangjuncong, DING Weilong, ZHANG Xinwei, LUN Yasen, GAO ruize

(School of Artificial Intelligence and Computer, North China University of Technology, Beijing 100144, China)

Abstract: Currently, there are still two issues with the assessment methods for competency attainment: ① Key competency elements rely on teachers' subjective construction, resulting in incomplete coverage and inconsistent standards; ② The evaluation perspective for competency attainment is singular, lacking fine-grained analysis of multiple teaching processes. To address this, an intelligent student competency attainment assessment method integrating large language models is proposed. First, a prompt template is designed for competency element extraction, leveraging large language models to automatically extract key competency elements from course syllabi, achieving both operational convenience and semantic consistency. Second, a "question-competency element" mapping and competency weighting model are constructed for the evaluation process, quantifying students' attainment of multi-dimensional competency elements through their responses. Comparing the model's calculated results with teacher evaluations reveals that the proposed method performs well in both accuracy and practicality, providing a feasible pathway for data-driven teaching assessment.

Key Words: large language models; students' ability achievement; teaching assessment; ability mapping and quantitative analysis; ability modeling

0 引言

在现代教育体系中, 达成性评价是衡量学生是否实现

课程目标的重要工具, 长期以来在基础教育和高等教育各阶段得到了广泛应用。传统达成性评价是以能力要素为导向, 教师基于学生期末成绩和预设能力集合两类数据进行评估^[1,2]。其中, 期末成绩源于期末考试或课程结束时

收稿日期: 2025-08-04

基金项目: 大学生创新创业训练计划(10805136025XN066-291)

作者简介: 刘阳俊聪(2001-), 男, 北方工业大学人工智能与计算机学院硕士研究生, 研究方向为大数据、智慧教育、人工智能等; 丁维龙(1983-), 男, 博士, 北方工业大学人工智能与计算机学院副教授, 研究方向为时空大数据、分布式系统、智慧城市; 张心玮(2004-), 男, 北方工业大学人工智能与计算机学院学生, 研究方向为软件开发、智慧教育等; 伦亚森(2001-), 男, 北方工业大学人工智能与计算机学院硕士研究生, 研究方向为大数据、智慧教育、人工智能等; 高瑞泽(2005-), 男, 北方工业大学人工智能与计算机学院学生, 研究方向为人工智能、智慧教育等。本文通讯作者: 丁维龙。

的综合测试,反映了学生对课程内容的整体掌握情况;预设能力集合则对应课程目标中的核心知识点或达成要求,教师通过手动分析学生的期末成绩,逐一比较成绩表现与能力集合,判断学生在哪些关键知识点或能力要素上达到了预期标准^[3]。例如,教师根据考试中特定题目的得分,推断学生是否满足某一能力要素,并据此形成评价结论。这种评价方式强调结果导向,操作相对简单,适合在大规模教育环境中实施,在实践中具有较高的普及性和制度稳定性。然而,传统达成性评价方法存在以下问题:

(1)传统达成性评价在能力要素提取方面效果不佳。该方法为了提升自动化程度,试图减少人工参与带来的主观性,对人工提取能力要素的方法进行了优化。例如,词频—逆文档频率(Term Frequency — Inverse Document Frequency, TF-IDF)等基于客观文本特征的能力要素提取技术,普遍缺乏对语义层面的深入理解,难以准确识别隐藏在课程大纲、学生培养计划等教学文本中的真实达成要求。同时,容易将无关内容的词语错误地纳入提取范围,导致整体提取效果不佳。特别是,在面向跨学科、多课程融合的复杂教学场景时,传统方法所构建的能力要素集合往往缺乏稳定的结构逻辑与语义一致性,限制了后续能力评价模型在实际教学场景中的复用性与标准化发展进程。此外,由于能力要素匹配的判断标准不统一,最终的评价结果可能无法真实反映学生的实际能力水平,进而导致“高分低能”等现象发生,揭示出评价内容与能力实质之间存在脱节。

(2)传统达成性评价方法视角单一,未对教学环节的多个过程进行细粒度分析。现有达成性评价方法主要依据单次测试结果对学生的课程目标达成情况作出判断,未持续观测学习过程中的能力发展轨迹和阶段性表现。此类结果导向的评价模式在一定程度简化了实施流程,但难以反映学生在实际学习中所体现的认知水平演进、问题解决策略、多维能力要素协同及非结构化任务表现等动态要素。这类能力在传统评价体系中往往被忽视,进而削弱了教学反馈的针对性与有效性。由此导致的后果是“低分高能”的现象,即部分学生在标准化测评中成绩不理想,却能在开放性或实践型任务中展现出较强的综合能力与创新潜力。因此,构建一种过程导向、动态可追踪且具有解释性的达成评价机制,对实现学生真实能力的精准刻画与课程教学质量的有效保障,具有重要的理论价值与实践意义。

本文研究基于北方工业大学教学实验评估与项目实践,主要贡献包括:①构建了一种基于大模型的题目—能力要素生成方法,有效缓解了传统方法在能力要素提取过程中存在的语义理解不足、提取范围划分不准确等问题;②提出基于所有作业/实验的判定机制,以每次实验/作业中的题目为基本单位,计算出总体达成度,从微观到宏观的视角评估学生能力,弥补了传统达成性评价视角单一、

粒度过粗的问题。

1 相关工作

1.1 能力要素及其提取方法

能力要素作为教学内容的载体,是连接课程设计与学生能力评估的桥梁。然而,当前主流方法在实际应用中面临诸多挑战,特别在抽取效率、语义一致性方面。

早期,能力要素抽取方法主要基于浅层文本特征,例如词频统计、TF-IDF等技术,依靠统计规律识别文本中的高频词或显著词^[4-6]。这类方法虽然实现简单、计算高效,但缺乏语义理解能力,因此难以准确区分具备“能力倾向”的动词型表达与常见的“知识性词汇”,从而限制了其在能力建模场景中的应用效果。为此,研究者引入更具语义表示能力的自然语言处理模型,例如基于图结构的无监督文本处理算法(TextRank)、基于主题建模的概率主题模型(LDA)、基于上下文语义理解的预训练语言模型(BERT),以增强能力要素抽取的上下文适应性与语义区分能力^[7,8]。尽管,上述方法在关键术语识别方面取得了较好效果,但大多数聚焦于词汇抽取本身,对能力要素与真实能力之间的语义对齐与结构对照问题关注较少,缺乏可用于教学实践的完整路径。

针对这一空白,智能教育研究逐渐将关注点从“能力识别”扩展至“关键词—知识点—能力维度”三者间的结构建模路径。歹杰等^[9]提出基于图神经网络的建模方法,将用户、课程与知识构建为异构图,通过图结构挖掘教学内容与能力要素之间的深层语义关系,为能力画像和精准推送提供支持。与此同时,能力维度本体的构建也成为了关键的研究方向。作为教育心理学中的经典理论,Bloom教学目标分类法将学生能力划分为认知、情感和动作技能3个领域,并进一步细化为多个认知层级,成为了当前教学目标设计与能力评价的重要基础^[10]。尽管该理论已被广泛应用于课程目标构建,但如何将课程文本中自然语言表述的教学目标自动映射至Bloom体系,仍缺乏通用化、系统化的技术工具。目前,较为先进的做法是借助能力本体构建和语义匹配模型,或利用BERT等深度语义模型实现教学目标与能力标签之间的自动对齐,但整体技术路径仍处于探索阶段,缺乏可大规模部署的稳定方案。

在能力识别与动态更新方面,深度知识追踪(Deep Knowledge Tracing, DKT)模型为学生能力的时序预测提供了有效思路^[11,12]。该模型通过建模学生作答序列,利用循环神经网络(Recurrent Neural Network, RNN)、Transformer等结构模拟学生对知识点掌握程度的动态演化情况,能较准确地预测其后续学习表现^[13]。然而,DKT模型依赖教师预先标注的知识点体系,自动建立能力要素提取与学生真实能力之间的映射能力较弱,尚不足以实现对能力要素的端到端建模与评估。

在数据采集与平台开发方面,已有研究进一步拓展了能力要素提取方法与建模场景。王宁等^[14]设计的智慧教育公共服务平台运行监测系统,通过监控运行日志和数据共享频率等指标,可得到学生访问学习站点的频率。王翔等^[15]在 AIGC 背景下构建的智慧教育平台,提倡以培养目标为导向构建课程资源与能力匹配机制,为学生提供定制化的学习资料与案例。唐利^[16]、林莉^[17]尝试在教育场景中引入人工智能教学助手,利用 AI 辅助教师开展教学工作或借助平台收集学生学习数据,通过深度学习分析学生在课堂中的表现与能力达成度。

然而,现有研究主要集中在特定教学场景,缺乏跨学科、系统化的能力提取方法。针对学生学习指标的构建,尚未构建一套自动化建模方法,仍依赖浅层文本特征分析及教师的主观语义理解与能力要素提取。尽管,部分研究已初步将课程内容与能力要求进行对齐,但在复杂、跨课程的实际教学环境中,仍存在能力要素识别不准确、能力匹配不统一、系统应用效果有限等问题。

1.2 学生能力达成性评价方法

能力达成度评价是教育质量保障体系中的核心环节,目标在于全面反映学生在教学过程中的能力发展水平。目前,主流的评价方法多聚焦于阶段性成绩或一次性考试结果,无法持续跟踪学生在多过程场景中能力,难以准确刻画学生真实的能力达成度,导致评价结果片面,针对性和解释力较差。

目前,能力评价方法主要可归纳为三类:①基于在线行为轨迹的建模方法,通过分析学生在学习平台上的点击记录、停留时间、视频观看行为等数据,预测其学习偏好与能力表现;②基于作业与测验数据的持续性建模,利用知识追踪或认知诊断等模型,动态刻画学生的知识掌握过程与能力发展趋势;③致力于融合多源数据,例如课堂活动、实验结果、作业表现、答题记录等,构建综合性的能力评价框架。例如,刘琛^[3]在研究中设计了基于贝叶斯理论的智能教学平台,通过整合过程数据与课程能力目标,可实时分析、分级判断学生在教学过程中的能力状态。在具体技术路径方面,知识追踪模型已成为支撑个性化能力评价的重要方法^[18]。

从早期的贝叶斯知识追踪模型(Bayesian Knowledge Tracking Model, BKT)到近年来的 DKT,再到引入结构信息的变种模型。研究者在不断提升模型对学生能力状态变化的预测精度与适应性。董文浩^[2]提出的基于贝叶斯推理的能力建模方法,在结合课程结构信息的基础上,对学生的解题表现进行能力层面分析,体现了较高的建模可解释性。高杨^[19]指出现有知识追踪方法大多停留在“题目—知识点”层级,对课堂教学、实验活动、项目任务等复杂教学环节中产生的数据利用不足,评价粒度有限^[20]。

此外,部分研究尝试利用历史高表现学生的学习路径作为对照,通过行为轨迹比较和反馈模拟,间接评估当前

学生的能力达成度。赵楠楠^[21]基于“雨课堂”平台构建大学英语混合教学评价体系,强调了过程性数据的重要性。卢宇等^[22]从资源生成、课堂互动、个性化学习、教学评价4个方面探讨 AI 驱动课堂评价的可行路径。苏洪雨等^[23]利用自然语言处理(Natural Language Processing, NLP)分析学生数学知识点的达成情况。刘聪^[24]提供了一种基于大规模数据存储与并行处理的思路,从海量数据中分析学生的达成性评价,以解决教育数据量大、计算复杂的问题。姜力^[25]指出学生的学习过程受多种因素共同影响,例如知识间的先修关系、任务难度差异及个体认知特征等,仅依赖静态评分难以全面反映学生能力的发展动态,因此需要借助动态建模方法对路径演进进行更细致地刻画。

综上,能力达成性评价在方法和数据利用层面仍存在不足,单过程数据源与静态评价方法难以支撑面向复杂教学场景的能力测评需求。未来,亟需增加多过程细粒度分析,发展基于动态过程模拟的能力建模技术,从而实现对学生能力达成过程的持续跟踪与真实反映。

2 本文方法

2.1 概述

图1为本文方法整体框架。

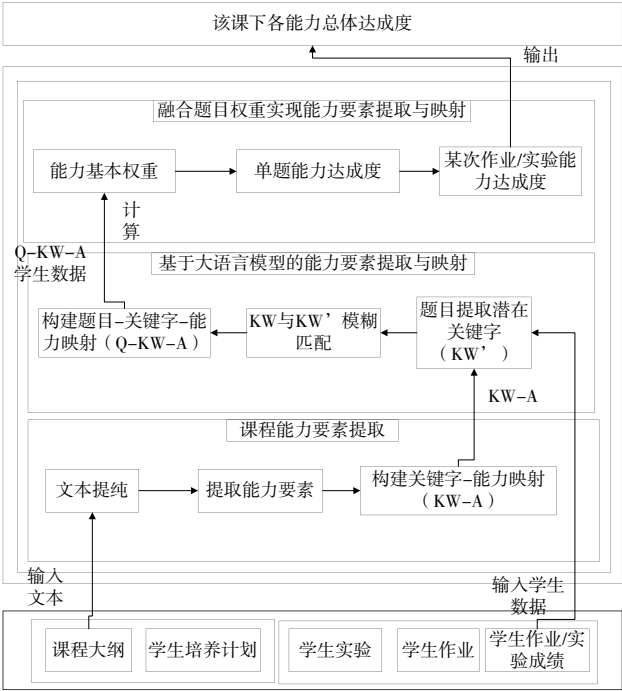


Fig. 1 Working framework of this article
图1 本文工作框架

研究主要聚焦两个方面:①在评价实施前,依据课程教学大纲与学生培养计划,基于大模型提取、量化与课程目标相符合的能力要素,并将其映射至具体的试题;②在评价过程中,基于题目所绑定的能力要素,结合学生作答情况评估其在相应能力要素上的达成度。能力总体达成

度用来量化学生对课程中各项能力完成程度。

由于关键字特征明显、易于提取,可通过教师手动构建,但能力要素十分抽象。为此,引入关键字(KW)作为中间变量,建立题目—关键字—能力要素(Q-KW-A)三元结构,得到题目—能力要素(Q-A)映射,为后续能力达成度分析提供数据支持。以数据库原理课程为例,详细阐述如何利用关键字对课程中的具体题目进行能力要素绑定,实现学生能力达成度的精细化评价与量化分析。

2.2 大语言模型驱动的能力要素提取与映射方法

为了实现课程能力要素的自动化提取与题目能力的精准映射,构建了一套基于大语言模型(Large Language Model, LLM)的课程关键能力要素提取流程,包括课程能力要素提取与题目能力绑定。

2.2.1 课程能力要素提取

首先,针对课程教学大纲、学生培养计划等文本资料进行结构化预处理,利用正则表达式对文本内容进行筛选

和清洗,剔除与课程能力目标无关的冗余信息,例如格式说明、空行、多余符号等非核心内容,保留描述课程知识点与预期能力达成目标的关键语义片段。该步骤旨在筛选出高度相关的课程目标与能力描述片段,为后续基于大语言模型的能力要素提取步骤,提供更精准、高质量的输入。其次,将处理后的文本输入大语言模型,结合设计的能力提取提示词,引导模型自动识别、输出具有代表性的课程能力要素集合。大语言模型能挖掘课程内容中蕴含的通用能力标签,通过设计提示词结构约束模型输出格式,可有效控制输出的可读性与标签规范性。最后,将能力要素集合交由教师来选择关键字与能力要素的对应关系,构造 KW-A 映射。通过人工构建关键字—能力(KW-A)映射表,不仅增强了标签体系的准确性,还为后续题目能力绑定提供了稳定的结构性基础。具体的提示词设计如表 1 所示。

Table 1 Ability element extraction prompt words

表 1 能力要素提取提示词

提示类型	示例
带有任务描述的基线提示	你是一位资深的高等教育专家,精通课程目标解析与能力培养要素的提取。你的任务是从给定的课程教学大纲、课程目标及人才培养方案文本中,识别出与课程核心知识点和能力培养密切相关的能力标签。请根据以下文本生成不同的层次,每个层次下有不同的能力要素,并以下列格式输出:
输出要求	1. 所有能力必须以“能力”结尾,例如“算法设计能力” 2. 每个能力短语不超过 8 个字 3. 层次名称严格按照 4 个字(包括层次) 4. 能力以层次格式输出 **xx 层次**: - xxx 能力 - xxx 能力 课程名称:数据库原理 课程性质:专业核心课程 课程目标:本课程旨在使学生掌握数据库系统的基本理论、设计方法及实际操作能力,能够理解关系数据库的基本概念、模式设计原则、查询语言 SQL 的使用、数据一致性与事务处理机制等内容。通过本课程的学习,学生将具备独立设计数据库系统的能力,能够分析实际应用需求并将其建模为数据库结构,同时掌握使用 SQL 语言进行复杂查询与数据操作的能力。课程还强调数据规范化、完整性约束、索引优化与并发控制等关键技术。
输入示例	
输出示例	**工程层次**:-工程问题分析能力;-工程方案设计能力 **技术层次**:-文献检索能力;-工具设备使用能力

2.2.2 题目能力自动绑定

在完成课程层面的 KW-A 构建后,将该结构应用于题目的能力标签识别任务。具体而言,对于课程中的实验题(以编程题为例),将题干输入大语言模型,以提取出每道题潜在关联的关键字集合,记为 KW'。相较于“直接给定关键字列表,让模型从中选择对应能力标签”的做法,更符合教师手动提取能力要素的过程,避免了“提示性”过强导致大模型产生错误联想的问题。随后,基于语义相似度计算方法,对 KW' 与教师已定义的关键字集合(KW)进行模

糊匹配,进而确定题目所涉及的能力项。换言之,将 KW、KW' 作为桥梁,构建能力要素与题目之间的联系。具体的提示词如表 2 所示。

2.3 基于题目权重融合的能力达成度评估方法

为了实现对学能力达成度的多粒度、全过程精准量化与评估,构建了一套基于学生所有作业和实验任务的能力要素评估体系。该体系对所有历史作业和实验中涉及能力要素的相关题目进行加权评分并聚合,以评估学生在特定能力上的长期学习成效,不仅融合了学生在单次作业

Table 2 Potential keyword extraction prompt words	
表 2 潜在关键字提取提示词	
提示类型	示例
带有任务描述的基线提示	你是一位资深的高等教育专家,精通课程目标解析与能力培养要素的提取。你的任务是从给定的题目与答案选项中提取最符合数据库原理这门课程的关键字
	请从以下数据库相关的题目和答案中提取 3~5 个最核心的关键字,要求:
输出要求	1. 必须是数据库领域的专业术语 2. 优先保留答案中出现的术语 3. 用中文逗号分隔 课程名称:数据库原理
输入示例	题干:创建国家表 Nation,包含属性有,国家编号 nationkey(整型,主键)、国家名称 name(字符型)、地区编号 regionkey(整型,外键)、备注 comment(字符型)
输出示例	提取关键字:表,外键,参照完整性

或实验中的短期表现,还体现了其在多次作业和实验中的长期学习趋势。

在聚合过程中,该体系保持了各题目对评估结果贡献度的动态公平性。以能力要素 j ($ability_j$)为例,总体达成度的具体计算过程如算法 1 所示。第 56~77 行代码展示了能力要素 j 的计算过程。简而言之,在考虑所有 N 次作业/实验的题目集合情况下,记录与能力要素 j 有关的所有作业/实验的得分与总分和,取其得分率。能力要素在整体层面的权重等于其基本权重除以 N 次作业/实验中绑定所有能力在其对应的作业/实验权重之和。最终总体达成度等于得分率乘以该能力在整体层面的权重。为了得到总体达成度,本文引入能力基本权重、单题上的能力达成度、单次作业/实验上的能力达成度 3 个变量进行计算。

(1)第 1~19 行代码计算能力基本权重。首先,统计所有关联能力要素 j 的题目;其次,计算这些题目得分率;最后,取其平均值作为能力要素 j 的能力权重。

(2)第 20~34 行代码计算单题上的能力达成度。首先,计算题目得分率;其次,根据第 1~19 行代码得到该题绑定能力要素的基本权重总和,并将能力要素 j 的基本权重除以总权重得到其在该题上的权重;最后,以该权重乘以得分率计算出能力要素 j 在该题上的达成度。

(3)第 35~55 行代码计算单次作业/实验上的能力达成度。首先,确定本次作业题目集合 Q_a ,并计算其中每道题的得分率;其次,筛选出与能力要素 j 绑定的题目,统计其总得分 $P_{a,j}$ 、总分 $M_{a,j}$;再次,计算该能力在本次作业中的相对权重,即其基本权重除以作业所涉及全部能力的基本权重之和;最后,将得分率与该能力在本次作业中的权重相乘,得到能力 j 在该次作业/实验上的达成度。通过对这门课中能力 j 下所有学生的能力达成度求平均值,即可得到该能力在这门课下的总体达成度。

综上,该指标体系的优点在于:①精度高,能细致区分

能力间差异;②公平性好,避免了题目结构复杂性或能力重叠性带来的失真;③适用于教学反馈与个性化学习策略制定,有助于教师精准施教,学生自我调节学习路径。

算法 1 能力达成度评估计算

输入:所有作业/实验集合(集合 AssignmentSet);学生成绩(成绩 ScoreData)。

输出:能力要素 j 总体达成度(达成度 finalAttainment $_j$)。

function finalAttainment $_j$ Calculation(集合,成绩)

==== 阶段一:计算能力基本权重 ====

```
1: for each ability $_j$  in 能力集合 do
2:   relatedQuestions $_j$  <- [绑定了能力要素 $j$ 的题目集合]
3:   for each assignment in 集合 do
4:     for each question  $q$  in assignment do
5:       if  $q$  绑定了能力  $ability_j$  then
6:         relatedQuestions $_j$ .add( $q$ )
7:       end if
8:     end for
9:   end for
10:  totalScore <- 0
11:  totalFullMark <- 0
12:  for each  $q$  in relatedQuestions $_j$  do
13:    totalScore += 学生在题目  $q$  上得分
14:    totalFullMark +=  $q$  的满分
15:  end for
16:  scoreRate $_j$  <- totalScore / totalFullMark
17:  abilityW $_j$  <- scoreRate $_j$  / relatedQuestions $_j$ .size()
18:  保存  $ability_j$  的基本权重  $abilityW_j$ 
19: end for
```

===== 阶段二:单题上的能力达成度 =====

```
20: for each assignment in 集合 do
21:   for each question  $q$  in assignment do
22:     scoreRate $_q$  <- 学生在该题得分 / 该题满分
23:     A $_q$  <-  $q$  绑定的所有能力集合
24:     sumWeight <- 0
25:     for each ability $_j$  in A $_q$  do
26:       sumWeight +=  $abilityW_j$ 
27:     end for
28:     for each ability $_j$  in A $_q$  do
29:       aj $_q$ _weight <-  $abilityW_j$  / sumWeight
30:       attainment $_{aj\_q}$  <- aj $_q$ _weight * scoreRate $_q$ 
31:       保存  $ability_j$  在题  $q$  上的加权达成度 attainment $_{aj\_q}$ 
32:     end for
33:   end for
34: end for
```

===== 阶段三:单次作业中能力的达成度 =====

```
35: for each assignment  $a$  in 集合 do
36:   for each ability $_j$  in 能力集合 do
37:     relatedQuestions <- []
38:     for each question  $q$  in assignment do
39:       if  $q$  绑定了  $ability_j$  then
40:         relatedQuestions.add( $q$ )
```

```

41: end if
42: end for
43: scoreSum <- 0
44: fullMarkSum <- 0
45: for each q in relatedQuestions do
46: scoreSum += 学生在题目 q 上得分
47: fullMarkSum += q 的满分
48: end for
49: R_aj <- scoreSum / fullMarkSum
50: sumAbilityW <- assignment 中所有绑定能力的基本权重之和
51: weight_aj <- abilityW_j / sumAbilityW
52: attainment_aj_assignment <- weight_aj * R_aj
53: 保存 ability_j 在 assignment a 中的达成度 attainment_aj_assignment
54: end for
55: end for
# ==== 阶段四: 计算能力的总体达成度 ====
56: for each ability_j in 能力集合 do
57: relatedQuestions <- []
58: for each assignment in 集合 do
59: for each question q in assignment do
60: if q 绑定了 ability_j then
61: relatedQuestions.add(q)
62: end if
63: end for
64: end for
65: scoreSum <- 0
66: fullMarkSum <- 0
67: for each q in relatedQuestions do
68: scoreSum += 学生在题目 q 上得分
69: fullMarkSum += q 的满分
70: end for
71: R_j <- scoreSum / fullMarkSum
72: sumAbilityW <- 所有作业中出现的的能力基本权重之和
73: weight_j <- abilityW_j / sumAbilityW
74: finalAttainment_j <- weight_j * R_j
75: 保存 ability_j 的总体达成度 达成度 = finalAttainment_j
76: end for
77: return finalAttainment_j
78: end function

```

3 实验结果与分析

3.1 实验环境与设置

硬件环境为: CPU 为六核心十二线程架构的 AMD Ryzen 5 4600H 3.0 GHz 处理器, GPU 为 NVIDIA GeForce GTX 1650 4 GB 显存, 可支持基础的深度学习模型训练与加速推理。软件环境为: Windows 操作系统, Python 版本为 3.13.3, Python 依赖管理工具为 pip 24.3.1, 所有实验依赖库

均通过 pip 安装于本地。在该操作环境下, 可确保代码在高版本 Python 下的兼容性与运行效率, 借助较新版本的 pip 管理工具, 也可提升依赖项的管理效率与环境的可维护性。

本文能力要素提取与映射实验采用的大语言模型为 GPT-4、Claude-3.5-Sonnet。其中, GPT-4 基于大规模预训练 Transformer 架构, 具备较强的文本生成与语义抽取性能^[26]; Claude-3.5-Sonnet 聚焦文本处理, 其架构在处理长上下文方面进行了重点优化^[27]。在实验过程中, 模型以 API 接口形式调用, 采用默认超参数配置, temperature 设定为 0.2, 以保证输出结果准确, top-p 参数保持 1.0, 最大生成长度为 2 000。

3.2 数据集与评价指标

本文数据集来自实际的课程实验数据, 涵盖 10 名学生在 5 次实验中的详细表现。每次实验由多个题目组成, 每道题记录了得分、满分、能力标签及其对应的达成度指标, 数据集结构完整、标签清晰, 支持从题目、实验到整体层面的多粒度学习表现分析。数据集信息如表 3 所示。

Table 3 Experimental dataset information

表 3 实验数据集信息

学生/人	实验/次	作业/次	能力/个	记录/条
10	30	20	559	400

为了验证大语言模型从课程教学大纲及人才培养计划(简称为: 文本)中自动生成的能力项, 是否在语义上与输入核心数据一致, 提出全文匹配相似度 $\alpha > A$ 和 Top-3 段落融合相似度 $\beta > B$ 进行评价。其中, 全文匹配相似度用于衡量大语言模型生成的每条能力项与整份文本核心语义的一致性; Top-3 段落融合相似度用于分析能力项是否能被文本中 3 段关键段落支持, 探索其语义来源是否是文本中的核心要素。

此外, 选取文本中的 30 段文本, 将传统的 TF-IDF 统计特征提取方法、基于语义向量的相似度提取方法 BERT 作为对照组, 以验证本文方法的合理性, 评价指标为 Precision、Recall 和 F1 值。在能力达成度评估的效果与分析部分, 引入教师主观评分结果作为对照基准, 将教师打分转换为标准化能力分布, 通过计算与模型评估能力分布之间的余弦相似度与 KL 散度, 以量化两者的一致性。

3.3 实验结果

3.3.1 能力要素提取效果与分析

表 4 为两种语义的匹配结果。本文结合教师在人工标注过程中对能力与文本语义相关性的主观判断, 发现当能力项与全文的匹配度超过 0.5, 与文中语义最相关的 3 段落匹配度(即 Top-3 段落融合匹配度)超过 0.7 时, 往往被认定为“来源明确”或“语义贴合”。因此, 将 $A=0.5$ 视为高全文匹配相似度, 将 $B=0.7$ 定义为高 Top-3 段落融合相似度。实验表明, 模型生成能力项与原文的平均全文匹配相似度为 0.53, 与 Top-3 段落融合后的平均匹配相似度为

0.71,整体分布较为集中,均值和方差均能体现出较强的语义相关性。其中,“理论掌握能力”“模型算法研究能力”等能力项的全文匹配度达到0.6以上,具有明显的语义对齐关系;Top-3融合匹配度显著高于全文匹配,说明该能力项是在多个相关段落的基础上综合形成,反映了大模型在高层抽象与跨语境提取方面的有效性。

Table 4 Two semantic matching similarity results

表 4 两种语义匹配相似度结果

能力项	全文匹配度	Top-3 融合匹配度
理论掌握能力	0.674	0.793
模型算法研究能力	0.683	0.801
人工智能创新应用能力	0.590	0.816
训练推理能力	0.721	0.778
应用实施能力	0.638	0.722
工程知识能力	0.476	0.758
问题分析能力	0.427	0.833
设计/开发解决方案能力	0.497	0.656
工程与社会能力	0.556	0.624
个人和团队能力	0.495	0.747
环境可持续发展能力	0.395	0.662
研究创新能力	0.495	0.747
项目管理能力	0.401	0.565
研究能力	0.357	0.409
协调和组织能力	0.557	0.721

值得注意的是,部分能力项在“全文匹配相似度”上虽未达到较高水平($\alpha < 0.5$),但在“Top-3 融合段落相似度”中却表现突出($\beta > 0.7$),例如“工程知识能力”与“问题分析能力”。这一现象表明,该类能力的语义支撑主要集中于培养计划文本中的少数关键段落,而非广泛分布于全文。此外,Top-3 融合策略整体相较于全文匹配略高,进一步验证了将多个局部语义信息融合进行能力映射十分合理且必要,佐证了所提能力提取方法在保持生成结构完整性的同时,具备了良好的语义可溯性,可为能力达成度评估提供强有力的文本依据。实验结果与对照组比较结果如表 5 所示。

由表 5 可知,TF-IDF 方法依赖词频权重,无法处理教育文本中的隐含语义,表现较差。BERT 引入上下文后虽提升了准确率,但难以对跨段落的语义进行整合。相较而言,GPT-4 在 Precision、Recall、F1 等 3 项指标上均优于传统

Table 5 Performance of three methods in ability extraction tasks

表 5 3 种方法在能力提取任务上的性能

方法	Precision	Recall	F1
TF-IDF	0.32	0.41	0.36
BERT	0.48	0.55	0.51
GPT-4	0.72	0.78	0.75
Claude-3.5-Sonnet	0.80	0.85	0.83

方法,仅次于 Claude-3.5-Sonnet。

3.3.2 能力达成度评估效果与分析

为了进一步分析融合题目权重实现能力达成度评估的有效性,将教师手动评估中的离散型能力等级数据与系统生成的连续型能力加权得分进行映射与整合,分别计算二者的余弦相似度、KL 散度指标,并以散点图形式呈现于同一坐标系中,以对两者之间关系进行直观地比较,如图 2 所示。其中,横轴为数据点索引(按能力维度合并后的顺序),左纵轴为相应的余弦相似度值,右纵轴为 KL 散度值。由此可见,几乎所有数据点的余弦相似度均接近于 1,说明学生与教师在能力评估向量之间的方向高度一致,在能力结构的分布趋势上具有极高的一致性;大部分数据点的 KL 散度值接近于 0,仅有少量点存在较高的偏离,说明学生与教师的能力评估在概率分布层面具有较高的一致性。特别的,KL 散度作为非对称信息熵度量,能有效刻画概率分布间的差异,低散度值进一步支持了二者评估之间的相似性结论。

综上,学生能力达成的自动化评估结果在整体分布趋势、概率结构上,均与教师主观评价一致,验证了本文方法在能力建模与度量方面的可行性与有效性。

3.4 案例分析

目前,融合大语言模型的学生能力达成度智能评估方法,已在北京市某高校的数据库原理课程中得到应用。该方法通过大模型提取相应能力要素,结合易于提取的关键词,成功建立了题目与能力要素间的映射关系。通过 5 次作业/实验数据便可得出学生能力达成度评价,如表 6 所示。该表精准勾勒出了每位学生在各项能力维度上的掌握情况,便于教师了解学生短板和优势,通过能力总体达成度协助教师优化课程设计方案。

Table 6 Assessment results of student ability achievement

表 6 学生能力达成度评估结果

学号	发散思维能力	抽象思维能力	概念识别能力	模型构建能力	知识记忆能力	软件能力	逻辑思维能力
21 102 020 209	0.146 40	0.105 50	0.695 30	0.571 00	0.637 30	0.255 70	0.102 70
22 101 130 107	0.166 90	0.126 20	0.869 70	0.258 00	0.623 90	0.443 80	0.243 70
22 101 130 110	0.268 50	0.272 60	0.635 70	0.343 30	0.731 20	0.287 40	0.361 30
22 101 130 133	0.246 60	0.388 20	0.754 50	0.438 90	0.810 00	0.435 80	0.422 60
22 101 130 118	0.339 50	0.285 40	0.712 90	0.462 10	0.517 00	0.578 50	0.304 60
22 101 130 120	0.132 50	0.191 50	0.653 40	0.572 40	0.764 60	0.392 00	0.276 10
22 105 050 121	0.271 40	0.299 60	0.595 50	0.666 40	0.606 40	0.751 70	0.309 00
22 151 080 101	0.273 80	0.288 50	0.695 90	0.521 10	0.844 10	0.425 90	0.350 70
22 151 080 102	0.457 70	0.116 00	0.568 60	0.669 90	0.652 70	0.656 20	0.234 90
22 151 080 104	0.460 60	0.344 30	0.754 80	0.453 00	0.885 50	0.467 90	0.233 90
能力总体达成度	0.276 39	0.241 78	0.693 63	0.493 31	0.707 27	0.469 49	0.283 95

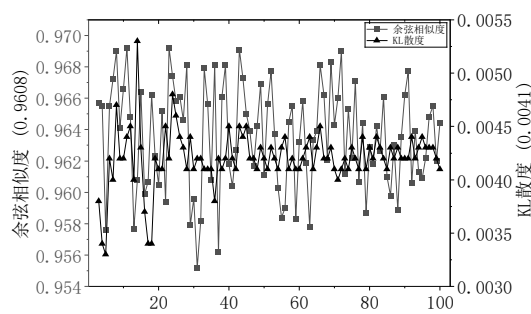


Fig. 2 Student-teacher evaluation distribution cosine similarity

图2 学生—教师评估分布余弦相似度

表7中所提能力项均为大模型在输入的教育文本中获得,通过对每位学生的学习情况进行多维度评估,便能通过学生视角了解该门课的达成度情况。由此可知,总体知识记忆能力和概念识别能力表现出色,发散思维能力、抽象思维能力和逻辑思维能力表现一般,证明数据库原理这门课的知识点对于学生而言十分抽象,难以理解,要真正掌握数据库底层原理并且延伸到数据库底层的优化过程,需要深入领会各概念间复杂的逻辑关系和内在联系,难度较大。

若以期末考试及格为最终目标,则仅需背诵记忆概念定义就足以满足基本要求,但该课程会考察学生能否根据设计模式正确设计数据库,或是否熟悉数据库语法。从模型构建能力和软件使用能力而言,学生的水平分布不均,在实操上存在明显差异:部分学生能熟练地将理论知识转化为数据库建模与查询实现,但部分学生在概念与实际操作中存在断层,缺乏对设计模式背后逻辑的深刻理解。这种差异不仅影响了他们在课堂任务中的表现,还反映出学生的数据库实践能力有待加强。

本文以22105050121这位学生为例,该名能力分布不均,强项在于软件能力、模型构建能力和知识记忆能力,说明在数据库工具的实际操作、编码和调试上较为熟练,可有效构建ER模型和数据库设计。其中,知识记忆能力指标说明该学生能较好地记住概念、公式和原理,有助于考试;概念识别能力中等,表示该学生能识别、区分关键概念,但需加强深度;发散思维能力、抽象思维能力和逻辑思维能力较弱则说明解决问题缺乏创意,具体问题抽象化能力较弱,例如无法从一个数据库问题联想到多种优化方案或从业务中抽象出数据库结构。为此,数据库原理课程的优化方向有两点:

(1)对于课堂。引入多样化教学,增加小组讨论,每周布置开放性题目,要求学生提交多种方案并讨论优缺点;教师提供真实案例供学生剖析其中的数据库结构,增加课堂中的证明题讲解,通过案例增强学生的逻辑思维能力。

(2)对于学生(22105050121)。增加逻辑推理练习,例如SQL查询优化题、关系代数证明等,通过伪代码或流程图分解复杂查询逻辑;多参加开源项目来提升自身软件工

程能力。

4 结语

本文针对当前教学评价中关键能力点依赖教师主观构建,耗时长、容易覆盖不全面和不一致的局限性,提出一种基于大语言模型的能力要素提取与映射方法。该方法的核心在于利用预设提示词模板,仅需引导大语言模型从课程大纲、培养计划中自动提取具备辨识度的能力点,解决了依赖教师人工审阅课程文件、提取关键字和匹配能力达成度的高工作负担与主观性问题。具体地,构建了一种融合题目权重实现能力达成度评估的计算方法,将学生答题表现映射至多维能力要素,并融合能力要素的基础权重,对教学环节多过程进行细粒度评估。

实验表明,所提方法输出的能力达成度与教师主观评估结果,在多个评价指标上保持高度一致,验证了其在实际教学场景中的可行性与有效性。相较于传统达成性评价,该方法能揭示学生在不同能力要素上的强弱分布,提升了评价的解释力与诊断价值。此外,该方法还具有良好的通用性与可拓展性,为推动精准教学、个性化学习提供了数据基础与技术路径。

参考文献:

- [1] XU J T. Research and design of a visual learning system based on knowledge space model [D]. Beijing: North China University of Technology, 2021.
徐锦堂. 基于知识空间模型的可视化学习系统研究与设计[D]. 北京: 北方工业大学, 2021.
- [2] DONG W H. Research on student ability evaluation and exercise recommendation algorithm based on Bayesian method [D]. Beijing: North China University of Technology, 2022.
董文浩. 基于贝叶斯方法的学生能力评价和试题推荐算法研究[D]. 北京: 北方工业大学, 2022.
- [3] LIU C. Design and development of an intelligent teaching platform based on bayesian student ability grading evaluation [D]. Beijing: North China University of Technology, 2024.
刘琛. 基于贝叶斯学生能力分级评价的智能教学平台设计与开发[D]. 北京: 北方工业大学, 2024.
- [4] NOMOTO T. Keyword extraction: a modern perspective [J]. SN Computer Science, 2023, 4(1): 92.
- [5] MORENO O A. Making sense of large social media corpora [M]. Berlin: Springer, 2024.
- [6] XU P Y, LIU H F, LIU B, et al. A survey of tag recommendation methods [J]. Journal of Software, 2022, 33(4): 1244-1266.
徐鹏宇, 刘华锋, 刘冰, 等. 标签推荐方法研究综述[J]. 软件学报, 2022, 33(4): 1244-1266.
- [7] CHEN K J, HUANG S Y. Improved RAKE algorithm for automatic keyword extraction from Chinese short texts [J]. Journal of Chinese Computer Systems, 2021, 42(6): 1171-1175.
陈可嘉, 黄思昱. 中文短文本自动关键词提取的改进RAKE算法[J]. 小型微型计算机系统, 2021, 42(6): 1171-1175.
- [8] RAJPUT S, GAHOI A, REDDY M. N-grams textrank a novel domain

- keyword extraction technique [EB/OL]. <https://aclanthology.org/2020.icon-termtraction.3.pdf>.
- [9] DAI J, LI Q S, CHU H, et al. Advancing smart education: course recommendation systems based on graph learning [J]. Journal of Software, 2022, 33(10): 3656-3672.
- 歹杰,李青山,褚华,等. 突破智慧教育:基于图学习的课程推荐系统[J]. 软件学报,2022,33(10):3656-3672.
- [10] XIE J L, HE F, ZHANG H G. Analysis of graduation requirements' connotation, logic, and teaching strategies from the perspective of bloom's taxonomy of objectives [J]. China University Teaching, 2025 (4): 44-52.
- 谢峻林,何峰,张红国. 基于BLOOM目标分类学视角的毕业要求内涵、逻辑、教学策略解析[J]. 中国大学教学,2025(4):44-52.
- [11] SU H, LIU X, YANG S, et al. Deep knowledge tracing with learning curves[J]. Frontiers in Psychology, 2023, 14: 1150329.
- [12] WANG Y, ZHU M X, YANG S H, et al. A survey and performance comparison of deep knowledge tracing models [J]. Journal of Software, 2023, 34(3): 1365-1395.
- 王宇,朱梦霞,杨尚辉,等. 深度知识追踪模型综述和性能比较[J]. 软件学报,2023,34(3):1365-1395.
- [13] LIU T Y, CHEN W, CHANG L, et al. Research progress on knowledge tracing based on deep learning [J]. Journal of Computer Research and Development, 2022, 59(1): 81-104.
- 刘铁国,陈威,常亮,等. 基于深度学习的知识追踪研究进展[J]. 计算机研究与发展,2022,59(1):81-104.
- [14] WANG N, XU J Y, CHEN X, et al. Design and practice of an operational monitoring system for online smart education platforms: a case study of the national smart education public service platform [J]. Modern Educational Technology, 2025, 35(7): 128-136.
- 王宁,徐竞祎,陈旭,等. 在线智慧教育平台运行监测系统的设计与实践——以国家智慧教育公共服务平台为例[J]. 现代教育技术,2025, 35(7):128-136.
- [15] WANG X, XIA L. Design of a smart education platform for software development talent training in the context of AIGC [J]. Industrial Control Computer, 2025, 38(6): 169-170, 172.
- 王翔,夏露. AIGC背景下软件开发人才培养智慧教育平台设计[J]. 工业控制计算机,2025,38(6):169-170,172.
- [16] TANG L. Application of artificial intelligence technology in preschool education teaching in higher vocational colleges [J]. Science & Education Guide, 2025(2): 17-19.
- 唐利. 人工智能技术在高职学前教育专业教学中的应用[J]. 科教导刊,2025(2):17-19.
- [17] LIN L. Application of deep learning in secondary school physical education based on smart platforms [J]. Fujian Sports Science and Technology, 2024, 43(6): 119-124.
- 林莉. 基于智慧平台的中学体育深度学习的应用研究[J]. 福建体育科技,2024,43(6):119-124.
- [18] JU S G, KANG R, ZHAO R M, et al. Deep knowledge tracing model with multi-concept fusion embedding [J]. Journal of Software, 2023, 34 (11): 5126-5142.
- 琚生根,康睿,赵容梅,等. 多知识点融合嵌入的深度知识追踪模型[J]. 软件学报,2023,34(11):5126-5142.
- [19] GAO Y. Research on dynamic knowledge tracing methods oriented to teaching process [D]. Beijing: North China University of Technology, 2023.
- 高杨. 面向教学过程的动态知识追踪方法研究[D]. 北京:北方工业大学,2023.
- [20] YUN Y, DAI H, ZHANG Y P, et al. A survey of personalized learning path recommendation [J]. Journal of Software, 2022, 33 (12): 4590-4615.
- 云岳,代欢,张育培,等. 个性化学习路径推荐综述[J]. 软件学报, 2022,33(12):4590-4615.
- [21] ZHAO N N. Construction of a formative assessment system for college english blended teaching supported by the rain classroom learning platform [J]. Science & Technology Vision, 2022(8): 106-109.
- 赵楠楠. 雨课堂学习平台支持下的大学英语混合教学形成性评价体系建构[J]. 科技视界,2022(8):106-109.
- [22] LU Y, TANG X Y. The structural levels and progressive pathways of generative artificial intelligence empowering classroom teaching [J]. E-Educational Research, 2025, 46(6): 75-82, 106.
- 卢宇,汤筱珂. 生成式人工智能赋能课堂教学的形态层级与进阶路径[J]. 电化教育研究,2025,46(6):75-82,106.
- [23] SU H Y, YE S, XU S H, et al. Research on a mathematics communication environment based on AI-enabled teaching assessment [J]. Journal of Mathematics Education, 2025, 34(3): 1-8.
- 苏洪雨,叶珊,许世红,等. 基于人工智能赋能教学评价的数学交流环境研究[J]. 数学教育学报,2025,34(3):1-8.
- [24] LIU C. Design and Implementation of an online education analytics platform based on Hadoop [D]. Shanghai: East China Normal University, 2023.
- 刘聪. 基于Hadoop的在线教育分析平台的设计与实现[D]. 上海:华东师范大学,2023.
- [25] JIANG L. Research on modeling and parameter identification of learning process [D]. Beijing: North China University of Technology, 2020.
- 姜力. 学习过程的建模与参数辨识研究[D]. 北京:北方工业大学, 2020.
- [26] SHI Y X, PENG D Z, LIAO W H, et al. Exploring OCR capabilities of GPT-4 (vision): a quantitative and in-depth evaluation [DB/OL]. <https://arxiv.org/abs/2310.16809>.
- [27] Anthropic. Introducing Claude 3.5 sonnet [EB/OL]. <https://www.anthropic.com/news/claude-3-5-sonnet>.

(责任编辑:刘嘉文)