大模型智能体评估教学实施新方法：使用教学设计达成度

A new method of teaching implementation for large-model agent evaluation: use of instructional design to achieve degree

任德博[1]，龙陶陶[1]，陈增照[1]，杜睿擎[1]
[1]华中师范大学人工智能教育学部
deboren@mails.ccnu.edu.cn

【摘要】 大模型智能体及思维链推理的发展解决了文本评估中的挑战。教学内容的一致性长期以来一直是教师评估的重要组成部分，单纯基于国家标准的一致性没有考虑到教师的个人发展，使用教学设计作为中介的条件干预教师教学内容优化成为了更加的方法。本研究用 deepseekR1 大模型取代人工评估，开发一种使用智能代理评估教学设计达成度的方法。教学设计文本可以识别的五个部分：教学目标、重点难点、教学过程、教学策略和评估。利用智能代理工具进行文本话语分析，并基于语义匹配和成就范围开发了一种新的计算方法，赋予权重(ω)，增强其准确性和稳定性。该模型通过这种方法进行了微调。结果显示，新手教师与专家教师在教学设计的不同部分达成度上存在显著差异。
【关键词】 教学设计；教学达成度；AI 智能体；教学内容；课程标准

Abstract: *The development of large model agents and chain-of-thought reasoning solves the challenges in text evaluation. The consistency of teaching content has long been an important part of teacher evaluation, and the consistency based solely on national standards does not take into account the personal development of teachers, and the use of instructional design as a mediating condition to intervene in the optimization of teachers' teaching content has become a more method. In this study, the deepseekR1 large model was used to replace the manual assessment, and a method was developed to evaluate the achievement of instructional design using intelligent agents. There are five parts that can be identified in an instructional design text: instructional objectives, key difficulties, instructional process, instructional strategies, and assessment. The intelligent agent tool was used for text utterance analysis, and a new calculation method was developed based on semantic matching and achievement range, which was assigned weight (ω) to enhance its accuracy and stability. The model is fine-tuned in this way. The results showed that there were significant differences between novice teachers and expert teachers in the achievement of different parts of instructional design.*
Keywords: instructional design, teaching achievement, AI agent, teaching content, curriculum standards

1.引言

组织和持续改进教学内容对于提高教学效果和学生学习至关重要，因为它们直接影响学生的表现。教学内容不仅包括所教授的知识，还包括内容的顺序、活动和其他相关元素（Gamoran et al., 1997）。教师负责安排这些元素，因此有效组织教学内容成为教学评估中的一个重要挑战（Porter, 1995）。先前的研究表明，基于主观经验的解释或模糊的标准在评估教学内容时可能导致教师脱离评估过程，从而削弱其有效性(Nägel et al., 2023；Schmidt et al., 2011)。中共中央、国务院印发了《教育强国建设规划纲要（2024—2035 年）》强调了高素质教师队伍建设和学生核心素养的提升，教学内容是核心素养的主要体现，是向学生传递核心素养的重要工具。

2022 年我国新引发基础教育课程核心标准，取消了三维教学目标，提倡学科核心素养的培养。教学设计作为教师教学的第一手工具，是教师结构化和优化内容的宝贵工具，它定义了教学中关键组成部分，如教学目标、重点和难点、教学策略、资源、过程（活动、案例分

析）和评估（Moore & Kearsley, 1996）。在建构主义理论中，教学设计要强调以学生为中心，不仅要求学生由外部刺激的被动接受者和知识的灌输对象转变为信息加工的主体、知识意义的主动建构者，强调环境和媒体的作用（何克抗，1997）；智能时代的教学设计，强调在课前课中课后促进学生的智能发展（刘邦奇，2016）。当前我国教师的教学设计承载了其对核心素养的理解、教学观、学生观，对其课程内容有很大的指导意义。

教学设计达成度代表了教学内容对于教学设计的完成情况，当前研究越来越关注将教学内容、核心素养和学生学习成果与教学设计达成度相匹配（Iaconangelo et al., 2022; Van et al., 2023）。有效的教学设计需要教师理解和正确实施（Purwaningsih et al., 2024），理解和实施设计计划对于新教师的专业成长至关重要（Lin et al., 2024；侯红玲等人，2019；邹吉权等人，2018），但存在人工评估成本高、效率低、标准不统一的特点。大语言模型的出现对教育领域产生了重大影响，尤其是在支持教师方面，其影响显而易见为了辅助教师成长增强教师教学实施的评估（Dai et al., 2024; Gill et al., 2024; Naidu, 2023），尤其是在评估对齐方面，是它能够应对定性评分的挑战。引入思路链（CoT）推理提供了一种新颖的解决方案（卢宇, 余京蕾 & 陈鹏鹤，2024），思路链使模型能够阐明解决问题的过程，提供了对人类任务解决方法的见解。

本团队基于大模型思维链开发教学设计达成度评估和建议工具，辅助教学内容评估。通过提供高质量的案例研究和现实的推理逻辑，不断优化现有模型，从而增强评分功能。

2.评估标准的建立和评估工具的开发

大语言模型通过将评分机制和尺度纳入模型，并根据专家评级进行调整，系统不断优化和更新提示工程，确保增加客观性和可操作性。在系统智能评估中测量达成度指数标准的确立是一个关键过程，Smith 等人（1992）强调其计算应系统地基于标准。本研究定义了达成度标准，以教学设计为基准评估每个标准是否反映在教学内容中。这些标准包括教学目标、关键和难点、教学策略（创新）、教学过程（活动、案例研究）以及教学评价（形成性和总结性）（Gagne et al., 2005; Weston et al., 1995）。我们的方法与 Porter 的人工评分方法不同（Porter et al，2011），将具体的数值标准纳入教学设计中，因此需要一个新的评价标准提供给大模型作为理解框架。我们将语义匹配与 Porter 的描述性方法相结合，将其分为五个层次，并根据语义的匹配度对每个层次给予了部分量化的评分标准，每条标准对应一个教学设计的部分。具体细节见下表 1。

表 1 大语言模型评分标准

评分	评分标准
完全达成（5）	1. 教学内容充分体现了教学目标。 2. 重点和难点用适当的时间和篇幅加以强调。 3. 教学策略和方法已得到充分实施。 4. 教学过程严格遵循预先设计的顺序。 5. 完全体现运用评价方法。
大部分达成（4）	1. 超过三分之二的教学目标被纳入到内容中。 2. 重点和难点得到一些强调，尽管它们的长度和时间分配没有完全平衡。 3. 超过三分之二的教学策略得到了实施。 4. 教学过程基本完成，但顺序和比例缺乏完全的一致性。 5. 评估方法为学生提供了及时的反馈，尽管与原始计划存在偏差。
适中达成（3）	1. 超过一半的教学目标得到满足。 2. 包含关键难点，但与一般知识点混编。

	3. 只有计划中的教学策略和方法的一半得到了实施。 4. 大部分的教学过程已经完成，尽管顺序缺乏清晰度。 5. 评估方法给一些学生提供了反馈，但缺乏对整个班级的全面评估。
部分达成（2）	1. 只有三分之一的教学目标得到满足。 2. 关键和难点包括在内，但其强调程度甚至低于一般知识点。 3. 只有三分之一的计划教学策略和方法已经实施。 4. 教学过程不完整，顺序混乱。 5. 仅提及评估方法，但未提供预期反馈。
没达成（1）	1. 教学目标在内容中没有体现出来。 2. 关键和难点完全省略。 3. 设计中的任何教学策略或方法都没有得到实施。 4. 教学过程完全偏离了原计划。 5. 教学过程中完全没有评估和反馈。

教学设计元素的重要性各不相同。Martin（2011）在一项关于教学设计元素对齐的研究中，将教学目标列为最重要，而教学材料则被列为最不重要。然而，教学设计原则和学校政策可能有所不同。例如，在以活动为基础的课堂中，教学活动被视为核心组成部分（Han et al., 2015)。因此，在计算教学设计与教学内容之间的对齐之前，引入了一个权重因子ω。本研究中重要性的水平是通过层次分析法（AHP）确定的，权重因学校和地区而异。我们收集了北京40位专家教师和学校管理人员的意见，他们对评分进行了0到10的打分。评分分为两轮进行，第二轮紧随第一轮之后（CI1 = 0.005; CI2 = 0.10）。最终结果使用了两轮的平均分数，显示教学目标的重要性最高，而教学评估的重要性最低（CI = 0.004)。评分结果及其对应的权重见表2。

表2APA权重评分结果

	教学目标	教学重难点	教学过程设计	教学策略	教学评价	ω
教学目标	1	1.435	1.3135	1.264	1.4235	.253
教学重难点	0.679	1	1.186	1.227	1.414	.207
教学过程设计	0.8945	1.075	1	1.3105	1.451	.221
教学策略	0.6645	0.7005	0.7755	1	1.1335	.161
教学评价	0.707	0.7075	0.6915	0.991	1	.158

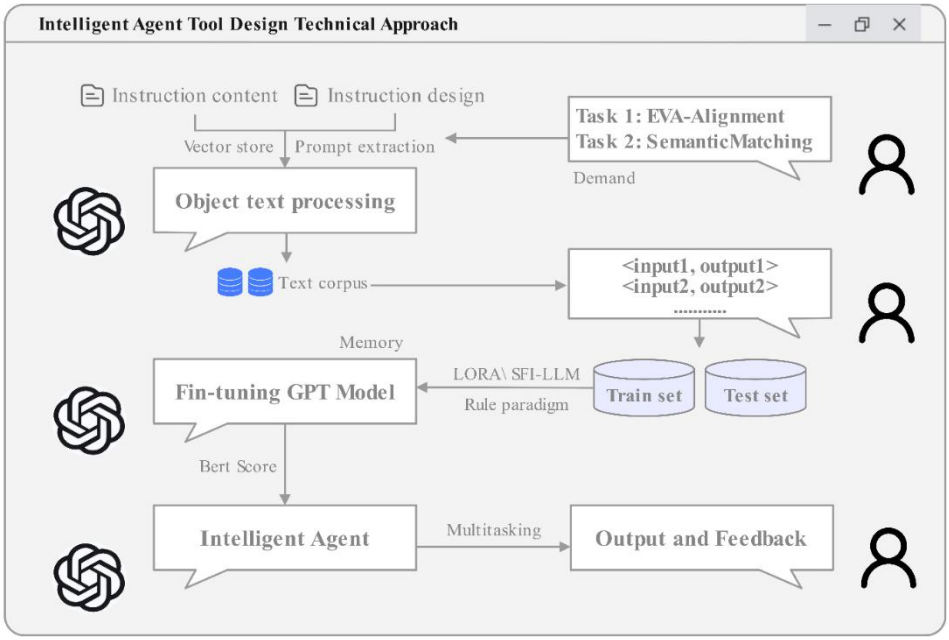

图 1 教学设计达成度测量的思维链设计过程

初始数据集是通过预先收集 45 节课的教学设计和内容创建的。由于这不是一个分类问题，因此需要进行模型预微调。我们采用了低秩适应（LoRA）方法（Mao et al., 2024），该方法有效地增加了可训练层，同时显著减少了 GPU 内存使用。预微调结果由 40 位教育学专家即行业老师通过层次分析进行了评估。专家们提供了改进建议，作为对比数据。我们比较了 DeepSeek 和 GPT 模型的性能。GPT 在强化学习和处理速度方面优于 DeepSeek，而 DeepSeek 在专业领域和复杂推理方面表现出色（Rahman et al., 2025; Gao et al., 2025）。我们选择了 deepseekR1 模型作为代理开发的基础模型。接下来，使用了 DeepSeek 模型构建了智能体推理。对于语义匹配，我们采用了 Bert Score 方法，该方法基于生成文本的标准参考指标评估语义相似度（Zhang et al，2019）。学习率设置为 3e-4，训练轮数设为 80。所采用的过程范式包括：规则范式、文本提取、单项目对齐评估(A)和语义匹配统计(G)。最终，使用这些方法获得了更准确的结果，如图 1 所示。

为了证明软件的有效性，对专家打分和智能体对教学设计达成度的打分进行了对比验证，通过数据看出专家打分的一致性明显低于智能体打分的结果。而专家打分的结果和智能体打分的结果具有较好的一致性，不具有显著性差异（p>0.05）。说明智能体评分在教学意义上拥有一定的有效性，如表 3 所示。

表 3 专家评分与智能体评分对比

	智能体评分		专家评分	
	Mean（STD）	IAA	Mean（STD）	IAA
教学目标	4.08（0.71）	0.836	4.16（0.83）	0.713
教学重难点	4.12（0.58）	0.871	3.97（0.71）	0.749
教学过程	3.84（0.71）	0.733	4.05（0.51）	0.581
教学策略	3.91（0.36）	0.786	0.63（0.58）	0.671
教学评价	4.01（0.63）	0.814	4.17（0.78）	0.813

** *一级指标一致性 (IAA)*

3.新手教师和专家教师教学设计达成度比较研究

以北京地区合作实验学校为依托，开展一项涉及 20 名中学教师的比较实验。在选定的教师中，5 名是新手，5 名是专家，且事先未通知参与者。教师来自两个科目：语文和数学，并在各组中均匀分布。从学期中期开始，每隔一周收集两组的教学设计和视频，直至为期五周的阶段测试。教学视频被转录并分析。目的是考察新手教师与专家教师之间的对齐差异。计算了 20 名教师的对齐度（M=6.399；SD=0.397）。实验中，我们发现新手教师与专家教师在教学方法上的排列没有显著差异；然而，当分解每个组成部分时，差异变得明显。所有课程均使用 Shapiro-Wilk 检验进行了正态性测试，结果显示负偏斜（-0.508<偏度<-0.043），表明接近正态分布。采用 t 检验来比较差异。结果显示新手教师与专家教师之间没有显著差异（t=-1.525；p=.145）；不同教育阶段之间也没有显著差异（t=-1.007；p=.327）；在教学目标（p = .211）、重点难点（p = .662）以及教学评价（p = .817）方面，新手教师与专家教师之间也没有显著差异。然而，在教学策略（p = .000）和教学过程（p= .032）方面发现了显著差异。在教学策略方面，新手教师的达成度显著高于专家教师（t = 5.689），而在教学过程方面，专家教师的达成度显著高于新手教师（t = -2.180）。我们进行了回归模型分析，以考察每个因素的影响，发现教学目标显著影响教学重点和教学过程的对齐度，但对教学策略和教学评价的达成度影响较小。教学重点和教学过程的达成度显著影响教学内容的对齐度。

研究发现，新手教师在教学实施方面面临挑战，尤其是在课程内容进展上，其一致性得分显著低于专家教师。教学策略专家型教师达成度低于新手型教师引发我们的兴趣，并进行关于原因的讨论，例如与年龄相关的教学习惯的影响或教师职业倦怠的效果。研究表明，年长的教师更容易出现职业倦怠（TSANG 等，2022），他们常常忽视教学策略的重要性，将其视为维护权威的形式工具。这往往是有害的，促使人们重新思考专家教师的定义。诸如丰富的奖项或多年的经验等标准常被用来区分专家教师，但这些标准并不全面。在以学生为中心的课堂中，促进学生学习的能力应该是区分新手教师和专家教师的关键因素。

4.教学启发和未来研究

教学设计与教学内容的一致性计算代表了一种新的方法，改进了 Porter 对教学标准的应用。作为指导教学的文件，教学设计满足了个性化教师评估的需求，并解决了不同学校之间缺乏统一评估标准的问题。在我国，大多数教学设计需经过上级部门的审查和批准，并需与国家 2022 年新印发的课程标准中核心素养保持一致，以确保教学评估的一致性，从而保证教学设计的合理性。当前强调以学生为中心课程时，标准化的教学标准通常会导致更高的评分；然而，这种方法忽略了教师的个人优势。这就是为什么我们使用教学设计作为标准来评估一致性，因为它有效地解决了这一问题，并且与学生的学习表现有很强的相关性。教学设计提供了更全面的内容覆盖，因此比国家标准，更方便教师达成度提升，而后者可能覆盖范围有限。同样，我们认为教学中的差异化是教学的关键原则，作为衡量标准，教学设计特别针对个别学生进行教学。其正确应用可以更准确地反映教学内容的质量。

智能体的构建有效减少了人类评估中的主观性。与其他智能评估方法一样，教研员同样在这一过程中起着关键作用，给出教师更具体的建议而不是完全机械化的评分。而智能体可以根据提出的框架提供多元化的建议，帮助教师选择性改进。未来改进框架的研究一定会作为最主要的研究方向。虽然仍有待改进的不足，但是本研究依然提出了一种创新性的方法帮助青年教师提升教学内容，说明大模型在教师教育、课堂管理、教师培训和学生学习中，尤其是话语分析和提升建议上，有着极大的潜力。进一步创建一个自动化的评估系统。该方法消除了手动评估的主观性，为不同课堂的教师提供了更有效的支持。其有效性已得到验证，突显了其应用的教育意义。它为未来教学培训的改革提供了新的参考。

参考文献

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. Educational Evaluation and Policy Analysis, 19(4), 325-338.

Porter, A. (1995). Research news and comment: The uses and misuses of opportunity-to-learn standards. Educational Researcher, 24(1), 21-27.

Nägel, L., Bleck, V., & Lipowsky, F. (2023). “Research findings and daily teaching practice are worlds apart” – Predictors and consequences of scepticism toward the relevance of scientific content for teaching practice. Teaching and Teacher Education, 121, 103911.

Schmidt, W. H., Cogan, L. S., Houang, R. T., & McKnight, C. C. (2011). Content coverage differences across districts/states: A persisting challenge for US education policy. American Journal of Education, 117(3), 399-427.

中共中央国务院印发《教育强国建设规划纲要(2024—2035 年)》,(2025-01-20).人民日报,006.doi:10.28655/n.cnki.nrmrb.2025.000833.

Moore, M. G., & Kearsley, G. (2012). Distance education: A systems view of online learning.

何克抗. (1997). 建构主义的教学模式、教学方法与教学设计. 北京师范大学学报(社会科学版), (05), 74-81.

刘邦奇. (2016). “互联网+”时代智慧课堂教学设计与实施策略研究. 中国电化教育, (10), 51-56+73.

Iaconangelo, C. J., Phelps, G., & Gitomer, D. H. (2022). Dimensionality and validity of the content knowledge for teaching construct using cognitive diagnostic modeling and known groups comparisons. Teaching and Teacher Education, 114, 103690.

Van Leent, L., Walsh, K., Moran, C., Hand, K., & French, S. (2023). Effectiveness of relationships and sex education: A systematic review of terminology, content, pedagogy, and outcomes. Educational Research Review, 39, 100527.

Purwaningsih, E., Sutoyo, S., & Suryadi, A. (2024). CoMCoRe-LS: an instructional design to enhance pedagogical content knowledge of pre-service physics teachers. Journal of Turkish Science Education, 21(2), 324-344.

Lin, X. F., Luo, G., Luo, S., Liu, J., Chan, K. K., Chen, H., ... & Li, Z. (2024). Promoting pre-service teachers’ learning performance and perceptions of inclusive education: An augmented reality-based training through learning by design approach. Teaching and Teacher Education, 148, 104661.

侯红玲, 任志贵, 何亚银 & 赵永强. (2019). 基于 OBE 理念反向设计教学过程研究. 大学教育, (10), 57-59.

邹吉权,刘晓梅 & 牟信妮.(2018).高职成果导向的教学设计与实施.中国职业技术教育,(20),26-32.

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., ... & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. Computers and Education: Artificial Intelligence, 7, 100299.

Gill, S. S., Xu, M., Patros, P., Wu, H., Kaur, R., Kaur, K., ... & Buyya, R. (2024). Transformative effects of ChatGPT on modern education: Emerging Era of AI Chatbots. Internet of Things and Cyber-Physical Systems, 4, 19-23.

Naidu, E. (2023). Leading Academics Believe Fears over ChatGPT Are Misplaced. University World News.

卢宇,余京蕾 & 陈鹏鹤.(2024).基于大模型的教学智能体构建与应用研究.中国电化教育,(07),99-108.

Gagne, R. M., Wager, W. W., Golas, K. C., Keller, J. M., & Russell, J. D. (2005). Principles of instructional design.

Weston, C., McAlpine, L., & Bordonaro, T. (1995). A model for understanding formative evaluation in instructional design. Educational technology research and development, 43(3), 29-48.

Porter, A., McMaken, J., Hwang, J., & Yang, R. (2011). Common core standards: The new US intended curriculum. Educational researcher, 40(3), 103-116.

Martin, F. (2011). Instructional design and the importance of instructional alignment. Community College Journal of Research and Practice, 35(12), 955-972.

Han, B., Gu, J., & Song, T. (2016). An Activity-based Instructional Design For Search Algorithm Expression of Elementary Students. Journal of The Korean Association of Information Education, 20(2), 161-170.

Mao, Y., Ge, Y., Fan, Y., Xu, W., Mi, Y., Hu, Z., & Gao, Y. (2025). A survey on lora of large language models. Frontiers of Computer Science, 19(7), 197605.

Rahman, A., Mahir, S. H., Tashrif, M. T. A., Aishi, A. A., Karim, M. A., Kundu, D., ... & Eidmum, M. D. (2025). Comparative Analysis Based on DeepSeek, ChatGPT, and Google Gemini: Features, Techniques, Performance, Future Prospects. *arXiv preprint arXiv:2503.04783*

Gao, T., Jin, J., Ke, Z. T., & Moryoussef, G. (2025). A Comparison of DeepSeek and Other LLMs. *arXiv preprint arXiv:2502.03688*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.