

2T-Agent: 生成式智能体驱动的教师思维 认知诊断框架

孙新杰^{1,2)} 刘淇¹⁾ 张凯¹⁾ 龚维印^{1,2)} 沈双宏³⁾ 汪飞¹⁾

¹⁾(中国科学技术大学认知智能全国重点实验室 合肥 230026)

²⁾(六盘水师范学院计算机科学学院 贵州 六盘水 553004)

³⁾(合肥综合性国家科学中心人工智能研究所 合肥 230051)

摘要 认知诊断在智能教育系统中日益重要,其核心目标是精准刻画学习者的知识状态与认知能力,以支持个性化教学。然而,现有方法大多依赖对答题表现的建模,忽视了学习过程中的认知动因、行为演化与情境因素,导致在复杂教学场景中适应性与解释力有限。为此,本文提出生成式智能体驱动的教师思维认知诊断框架(2T-Agent)。该框架模拟人类教师在认知判断过程中的三类核心能力:(1)设计任务感知的信念建模模块,整合题目结构、历史行为与上下文信息,生成情境化的认知信念;(2)构建具备观察与反思能力的认知行为模块,基于学生行为轨迹动态调整评估策略,并通过纠偏与总结反思机制实现认知修正与持续优化;(3)引入因果解释模块,结合大语言模型生成技术,自动构建“知识状态—行为过程—任务表现”的解释链条,提升模型的可解释性与干预价值。在多个认知评估基准数据集上的实验结果表明,2T-Agent 相较于现有神经与统计类 CD 模型,平均在准确率(ACC)上提高 1.61%,在 F1-Score 上提高 2.94%。特别在复杂语境下表现出更强的适应性与推理能力,展示了面向未来智能教育系统的广阔应用前景。为推动该领域的进一步发展,数据和代码已发布在 <https://github.com/xinjiesun-ustc/2T-Agent>。

关键词 教育数据挖掘; 认知诊断; 大语言模型; 智能体; 个性化学习; 可解释性

中图法分类号 TP18 DOI号 10.11897/SP.J.1016.2026.00574

2T-Agent: A Generative Agent-Driven Framework for Teacher-Like Thinking in Cognitive Diagnosis

SUN Xin-Jie^{1,2)} LIU Qi¹⁾ ZHANG Kai¹⁾ GONG Wei-Yin^{1,2)} SHEN Shuang-Hong³⁾ WANG Fei¹⁾

¹⁾(State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei 230026)

²⁾(School of Computer Science, Liupanshui Normal University, Liupanshui, Guizhou 553004)

³⁾(Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230051)

Abstract Cognitive diagnosis (CD) has become a fundamental component of intelligent education systems, aiming to accurately characterize learners' knowledge state and cognitive abilities for adaptive

收稿日期: 2025-07-07; 在线发布日期: 2025-10-23。本课题得到国家自然科学基金地区科学基金项目(62567004)、国家自然科学基金青年科学基金项目A类(62525606)、国家自然科学基金青年科学基金项目(62406303)、贵州省科技计划项目(黔科合基础MS[2026]753、黔科合基础MS[2026]754)、中国科学技术大学-六盘水师范学院对口合作发展联合基金(USTC-LPSNU-2025-10)、贵州省科技计划项目青年引导项目(黔科合基础-[2024]青年012)、贵州省高等学校本科教学内容和课程体系改革项目(GZJG2024331、GZJG2024323)资助。孙新杰, 博士研究生, 主要研究领域为智能教育和自适应学习。E-mail: xinjiesun@mail.ustc.edu.cn。刘淇(通信作者), 博士, 教授, 主要研究领域为数据挖掘。E-mail: qiliuql@ustc.edu.cn。张凯, 博士, 副研究员, 主要研究领域为大模型优化。龚维印, 博士研究生, 主要研究领域为多模态意图识别。沈双宏, 博士, 副研究员, 主要研究领域为智能教育。汪飞, 博士, 博士后, 主要研究领域为认知诊断。

and personalized instruction. By offering fine-grained insights into how learners understand, reason, and apply knowledge, CD supports data-driven educational decision-making. However, most existing CD methods primarily rely on modeling learners' answer performance while neglecting the underlying cognitive mechanisms that drive behavioral evolution and contextual adaptation. As a result, these models often struggle to maintain interpretability and robustness in complex learning environments involving multi-step reasoning, reflective behaviors, and diverse contextual conditions. Moreover, their reliance on static representations of student ability limits the capacity to capture dynamic cognitive change and individualized learning trajectories, which are critical for real-world educational applications. To address these challenges, we propose a Generative Agent-Driven Framework for Teacher-Like Thinking in Cognitive Diagnosis (2T-Agent). Inspired by the evaluative, reflective, and explanatory processes of human teachers, 2T-Agent introduces a generative paradigm that emulates teacher-like cognition for more interpretable and adaptive diagnostic reasoning. The framework incorporates three interacting modules that jointly realize dynamic assessment and reasoning. (1) The Belief Modeling Module performs task-aware perception by integrating question structure, historical learning trajectories, and contextual factors to generate contextualized cognitive beliefs, which dynamically represent learners' latent cognitive state and support task-specific inference. (2) The Cognitive Behavior Module incorporates observation and reflection mechanisms to adaptively adjust diagnostic strategies according to learners' behavioral trajectories. It applies corrective reflection to revise inconsistent estimations and summary reflection to consolidate accumulated knowledge, thereby achieving continuous optimization of assessment accuracy. (3) The Causal Explanation Module leverages large language model (LLM)-based generation to automatically construct causal explanatory chains that connect knowledge state, behavioral processes, and task performances. These explanations enhance interpretability and provide actionable feedback for targeted educational interventions, bridging the gap between algorithmic reasoning and human pedagogy. Extensive experiments were conducted on multiple cognitive assessment benchmarks to evaluate the effectiveness and generalization of 2T-Agent. The results show that 2T-Agent achieves an average improvement of 1.61% in accuracy (ACC) and 2.94% in F1-Score compared with existing neural and statistical CD models. Furthermore, it demonstrates stronger adaptability and reasoning ability in complex contextual scenarios, indicating robustness under dynamic and heterogeneous learning conditions. Ablation studies verify that the belief modeling and reflective behavior modules are key contributors to performance improvement, while the causal explanation module enhances interpretability without increasing computational cost. Qualitative analysis further reveals that 2T-Agent can emulate teacher-like reasoning, producing coherent and logically consistent explanations that align with human evaluative logic. To facilitate further research in this field, the data and code have been made publicly available at <https://github.com/xinjiesun-ustc/2T-Agent>.

Key words education data mining; cognitive diagnosis; large language model; intelligent agent; personalized learning; interpretability

1 引言

随着教育数字化与智能化进程的不断推进, 认知诊断(Cognitive Diagnosis, CD)在教育场景中的地位愈发重要^[1-2]。其核心目标是准确判定学习者的

知识掌握状态与认知能力变化, 以支撑个性化教学决策与高质量反馈服务^[3]。相比传统的考试得分方式, 认知评估强调对学习过程、思维策略与能力表现的深入建模, 是实现精准教学和智能反馈的重要基础^[4-5]。

现有的认知诊断方法主要基于动态认知诊断又名知识追踪(Knowledge Tracing, KT)^[6]与静态认知诊断(CD)技术^[7],代表性方法包括项目反应理论(IRT)^[8]、深度知识追踪(DKT)^[9]和神经认知诊断模型(NCDM)^[10]。这些方法通过分析学生的历史答题数据,在标准化任务或封闭题库环境中取得了良好的表现。然而,真实教学情境远比实验环境复杂,学习者的行为具有显著的个体差异^[11]、策略变化^[12]与情绪波动^[13]特征,使得上述模型在泛化能力、认知刻画深度以及结果可解释性方面存在明显不足^[14-15]。

在智能教育系统中,精准建模学生的知识状态与识别潜在的认知偏差,是实现高质量个性化教学的关键基础^[16]。然而,现有认知评估方法多侧重于基于答题表现的预测精度,忽视了学习行为背后的认知动因与推理机制。这种“以表现代认知”的范式在面对任务异质性、策略多样性和情境波动等真实教学因素时,往往表现出适应性差、解释力弱的局限性。如图 1 所示,现有方法缺乏结构化建模,难以揭示“认知状态-行为路径-任务表现”之间的内在关联。因此,认知诊断系统亟需从静态的标签式诊断转向面向过程的动态建模。

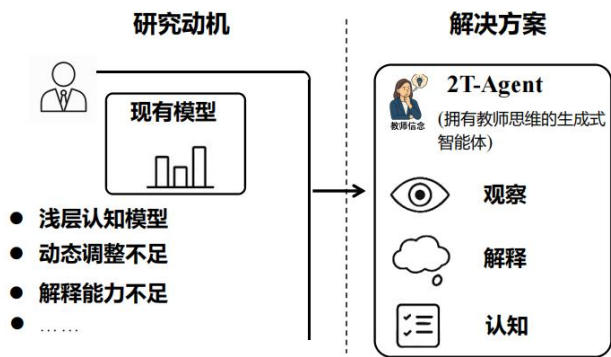


图 1 从现有模型到类教师思维智能体的概念跃迁:增强可观测性、可解释性与认知洞察力

为实现上述目标,智能教育诊断系统需应对以下关键挑战:第一,认知状态的情境依赖建模能力仍显不足。现有方法往往将认知状态简化为与任务无关的静态表示,忽略其对题目结构、知识点组合和任务情境的依赖性。在复杂语境中,如何基于有限的行为信息准确推理学生的真实认知状态,仍是一个尚未充分解决的问题。第二,缺乏对行为过程的建模与策略调整机制。学生的知识状态与行为选择具有显著的时间相关性,单次答题行为难以全面

反映其认知能力,模型亟需构建能够覆盖短期行为偏差与长期学习轨迹的行为记忆机制,并支持策略性反思与持续性优化。第三,现有方法缺乏结构化、因果性的认知解释能力。高质量的认知诊断不仅需要输出预测标签,更应能够解释其成因与演化路径,从而支撑有针对性的教学干预。然而,大多数方法尚不具备将认知表现转化为“知识掌握-策略选择-任务结果”因果链的能力,导致解释性不足、透明度较低。

在真实教学中,教师的认知评估行为体现出更高阶的认知能力。他们不仅能够基于有限信息判断学生的知识状态,还能结合长期观察、当前表现与教学经验,动态调整策略并推理学生的思维路径。为应对上述挑战,本文提出 2T-Agent 框架,一个面向认知诊断任务、具备类教师思维认知机制的智能体。该框架模拟教师在认知判断过程中的三类核心能力,具体包括:首先,教师信念建模模块,在复杂语境中整合题目结构、学生历史行为与上下文信息,生成关于学生知识状态的初步认知信念,实现面向具体任务情境的认知状态判断,为后续的动态调整与教学干预提供支持。其次,教师观察模块,结合学生的短期答题表现与长期行为轨迹,实现对认知状态演化过程的动态追踪与策略适应。一方面记录学生在当前任务中的局部偏差,另一方面构建其跨知识的认知演化路径。在此基础上,系统引入两级反思机制:纠偏式反思(Corrective Reflection)用于修正当前预测误差,总结式反思(Summary Reflection)用于周期性优化策略参数,模拟教师“观察-反思-调整”的认知过程。最后,解释与认知链生成模块面向可解释性建模,通过追踪学生的行为路径与答题序列,结合语言模型生成机制,自动构建“知识状态-行为过程-任务表现”之间的因果推理链条。该机制输出结构化解释,揭示认知偏差的潜在成因,并提供具备教学干预价值的反馈建议,赋予模型类教师的解释与诊断能力。

我们的主要贡献总结如下:

(1) 提出 2T-Agent,一个具备类教师认知机制的智能体框架,系统建模教师在认知判断中的三类核心能力。

(2) 设计任务感知的信念建模模块,使框架具备人类教师般的判断直觉与情境理解能力。

(3)实验结果表明,2T-Agent在多个认知评估任务中显著优于现有主流模型,特别在解释性与情境适应方面表现出色。

2 相关工作

准确评估学习者在学习阶段的认知状态,并基于此,评估其在面对不同题目时的真实表现状态,毫无疑问是一项极具价值的任务。接下来,我们将分别从认知诊断与基于大语言模型的生成式智能体两个角度,分别介绍相关的研究成果。

2.1 认知诊断

认知诊断(Cognitive Diagnosis, CD)旨在基于学生对试题的反应,推断其对知识概念的掌握情况,广泛应用于个性化教学与能力评估场景中^[17-18]。经典模型如IRT^[8]、MIRT^[19]和DINA^[20]通过建模学生能力与题目属性之间的关系,实现了对学生认知特征的有效刻画。这类方法通常适用于静态测试,如考试或阶段性评估,强调诊断结果的可解释性与稳定性。为弥补传统认知诊断方法在精度和细粒度建模方面的局限,神经认知诊断模型(Neural Cognitive Diagnosis Models, NCDM)已成为认知诊断研究的主流路径。相比传统方法,NCDM通过引入神经网络结构与多维参数之间的非线性交互机制,在提升诊断精度的同时,也增强了模型的表达能力与可解释性^[10]。围绕不同任务需求,学者们提出了一系列具有代表性的模型:如DCD模型设计了解耦与对齐机制,以精准建模学生能力、题目难度与标签分布之间的关系^[21];ACD模型将情绪感知融入认知状态建模,拓宽了诊断的视角与维度^[22];FairCD框架则通过能力分解策略,有效提升了模型对公平性与偏差的控制能力^[23]。此外,随着图神经网络(Graph Neural Networks, GNN)的发展,研究者进一步探索其在认知诊断中的应用,以建模学生、题目与知识点之间复杂的图结构关系^[24]。例如,RCD模型引入多轮特征学习机制,显著增强了模型的表示能力与诊断性能^[14]。

与此同时,动态认知诊断又名知识追踪(Knowledge Tracing, KT)致力于捕捉学生在连续学习过程中的知识状态演化,已广泛应用于智能教学系统与在线学习平台^[13,25]。自DKT^[9]等深度模型提出以来,KT在建模学生行为序列与预测未来表现方面取得了显著进展,强调时间序列建模与实时适应能

力。近年来,随着模型结构与诊断算法的持续优化,知识追踪模型在预测性能方面取得了显著进展^[26-27]。例如,THKT模型^[25]通过分层建模学习者的知识状态,利用教育学中的布鲁姆层次,有效提升了预测的准确性,也增强了模型的可解释性。DASKT模型^[13]则引入学习过程中情绪状态的建模,在有效模拟学生练习行为的基础上,进一步提升了诊断的准确性与可解释能力。

需要指出的是,CD与KT并非相互替代的技术路径,而是面向不同应用场景、目标互补的两类研究方向^[28-29]。前者强调在特定时间点的认知状态识别,注重解释性与细粒度刻画^[30];后者关注长期学习过程中的状态转变,注重预测性与动态建模能力^[31]。两者近年来在教育领域同步发展,相辅相成。本研究提出的2T-Agent借鉴认知诊断框架,构建了动态分析学生知识状态的教师智能体。不同于传统CD或KT模型,2T-Agent融合诊断的解释性与追踪的适应性,模拟教师持续分析学生理解路径与学习状态,实现更具专家特质的教育模拟与反馈。

2.2 基于大语言模型的智能体

近年来,以Generative Agents^[32]为代表的、由大语言模型(Large Language Models, LLMs)驱动的智能体(Agents)研究取得了突破性进展,并在教育领域展现出广泛的应用前景^[33-34]。这些智能体不仅能够模拟学生的行为,还能够辅助个性化教学、评估学习成效,并促进对学生认知过程的建模。其核心能力主要依赖于LLMs在上下文学习、记忆机制、工具调用和多步推理等方面的优势^[35-39]。

在学生行为建模方面,已有研究充分利用LLMs强大的语言理解与生成能力^[3,40],通过设计具备交互性和响应性的代理,模拟学生在不同学习任务中的表现。例如,Edu4Agent框架引入了基于LLM的代理以模拟学生对教学内容的响应,从而为教育数据生成和训练提供了新范式^[41]。刘等人^[42]则尝试通过LLM预测学生在编程任务中的代码提交行为,尽管实现了行为的初步模拟,但粒度较粗,且缺乏对学习过程的连续建模。相比之下,SocraticLM引入启发式问答机制,以类教师角色引导学生完成推理过程,增强了教学场景中的交互性与思维启发^[43];EduAgent则通过模拟学生对多模态学习材料(如PowerPoint演示和视频)的学习行为来实现认知追踪,尽管具备一定效果,但其过度依赖人工预设的认知因子,限制了模型的通用性^[44]。

在支持个性化学习方面, LLM Agents 逐渐展现出路径规划与自适应推荐的能力, 能够根据学生的个体差异动态调整学习策略^[45-47]。相关研究引入长期记忆机制记录学生的学习轨迹, 同时结合短期记忆模块对当前任务进行即时响应^[48-49]。此外, 工具增强机制(tool augmentation)的引入进一步拓展了 LLM Agents 的能力边界, 例如结合知识追踪模型预测学生知识状态, 实现针对性学习内容推荐^[50]。工具增强能力也使得 LLM Agents 在更复杂的教育任务中展现出较强的泛化性与可扩展性。例如, Park 等人^[51]提出的 Generative Agents 能够模拟虚拟人类的日常行为, 包括记忆形成、行为执行与反思过程, 并被后续研究拓展至教育场景中, 用于建模

学生在学习任务中的认知变化与行为轨迹^[12,52-53]。这些工作表明, 具备长期记忆与工具协同能力的 LLM Agents, 能够胜任复杂的行为建模与路径规划任务。此外, 智能交互作为教育代理系统的核心能力之一, 也被广泛研究。通过持续交互与动态调整, LLM Agents 可以根据学生当前的知识水平和学习状态调整教学策略, 生成个性化反馈, 进而优化学习路径和提高学习效果^[54-55]。这种“以学生为中心”的代理行为模式, 为构建具备人类教师特征的智能教育系统提供了理论与实践基础。与现有教育智能体主要模拟学生行为的做法不同, 2T-Agent 的核心创新在于模拟教师的思维模式, 从而实现对学习状态认知状态的精准诊断。

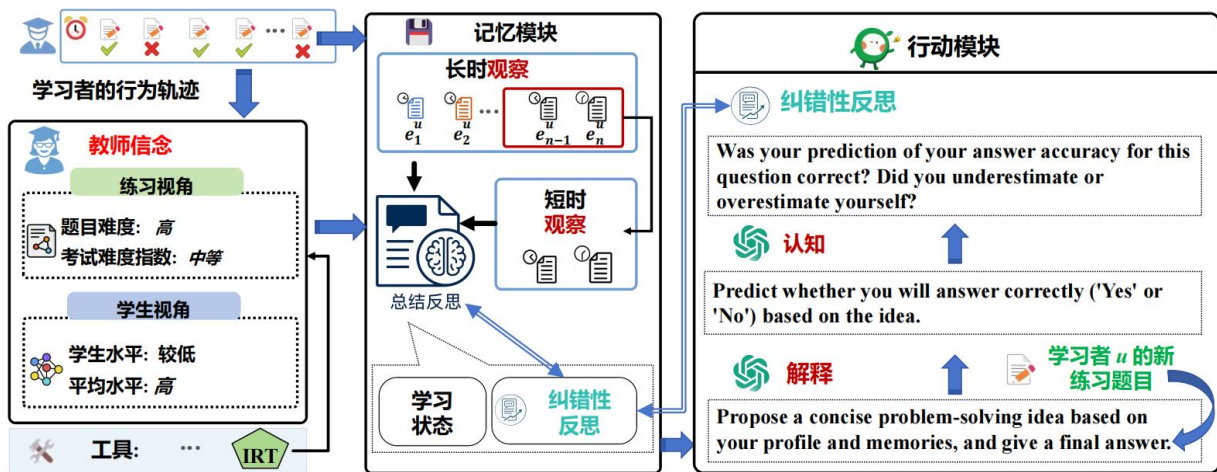


图2 2T-Agent 的整体框架: 一个由大语言模型驱动的智能体 $Agent_u$

3 准备工作

在第3节中, 我们详细阐述 LLM 驱动的智能体的基本概念, 并进一步介绍了 2T-Agent 框架中的关键组成部分。表1提供了本文所涉及的所有符号的汇总。

表1 符号和说明

符号	说明
AT, E	教师智能体集合与题目集合
KC, R	知识概念集合与答题结果集合
$Agent_t$	对应教师 t 的智能体
n, m, k	教师代理数量、题目数量、知识点数量
$B_{t,e}$	教师对学生 t 在题目 e 上掌握情况的信念
M_u, M_u^{infer}	教师代理对学生 u 的通用记忆模块和反思记忆模块
\hat{r}_t, r_t, δ_t	教师预测结果、学生真实作答结果与预测误差
CR_t, SR_t	教师代理生成的纠偏反馈与阶段总结
π_t, a_t	预测的解题策略与模拟作答内容

3.1 问题定义

在 2T-Agent 框架中, 我们引入若干核心集合以支持对学习者的认知评估的建模过程: 学习者专属教师智能体集合 $AT = \{a'_1, a'_2, \dots, a'_n\}$, 包含 n 个智能体, 每个智能体对应一位特定学习者, 模拟其专属教师角色; 练习集合 $E = \{e_1, e_2, \dots, e_m\}$, 由 m 道练习题构成; 知识概念集合 $KC = \{kc_1, kc_2, \dots, kc_k\}$, 涵盖 k 个目标知识概念; 以及作答响应集合 $R = \{0, 1\}$, 用于表示学习者对题目的作答结果, 其中 1 表示答对, 0 表示答错。基于这些集合, 我们将智能体的评估行为建模为三元组 (a', e, r) , 其中 $a' \in AT$ 表示特定教师智能体, $e \in E$ 表示对应练习题, r 表示学习者的作答结果(ground truth)。该建模方式为构建像教师一样思考的智能体提供了数据基础。

生成式智能体驱动的教师思维认知诊断框架(2T-Agent): 给定学习者的测试记录 $L_s = \{(a'_t,$

$e_1, r_1), (a_i^t, e_2, r_2), \dots, (a_i^t, e_t, r_t)\}$, 我们的目标是模拟学习者专属教师的思维模式来评估学习者。同时, 通过引入基于大模型驱动的智能体, 确保评估结果更为贴近学习者的真实表现并兼具可解释性。

4 2T-Agent 框架

在 2T-Agent 中, 智能体以大语言模型(LLM)为核心架构, 模拟教师的思维模式^[56], 具备观察、解释与认知的能力^[57], 从而实现对学习者的全面理解。图 2 展示了 2T-Agent 框架的整体架构, 而算法 1 给出了其详细的流程步骤。为实现类教师化的精准评估, 我们为每位学习者 s 构建一个专属教师智能体 $agent_s$, 实现一对一的认知交互。每个智能体内置一个教师信念模块, 用于建模学习者的个性化行为模式与习题属性。同时, 智能体还包含一个记忆模块, 用于观察学习者的历史纪录与近期表现, 并对其学习过程进行总结性与纠正性反思, 以构建更具抽象层次的认知表征。此外, 为了更自然、连贯地模拟学习者的学习与认知行为, 教师智能体还集成了一个行动模块, 能够解释学习者的认知状态, 并据此推理出符合其个体特征的认知机制。

算法 1. 2T-Agent 框架

输入: 估计的学生能力 $\hat{\theta}_i$; 估计的问题难度 \hat{d}_e ; 历史行为序列 $H_u^{(1:t-1)}$; 当前任务 e_t ; 作答结果 r_t 。

输出: 解题策略 π_t ; 作答内容 a_t ; 预测的作答结果 \hat{r}_t' 。

1. 通过式(1)到(3)初始化教师信念 $B_{t,e}$;
2. 通过式(4)提取长期观察特征 $H_u^{(1:t-1)}$;
3. 通过式(5)提取短期观察特征 $H_u^{(t-w:t-1)}$;
4. 通过式(6)计算预测偏差 δ_t ;
5. 通过式(7)生成纠偏反馈 CR_t ;
6. 通过式(8)形成阶段性总结 SR_t ;
7. 通过式(9)更新长期反思记忆池 $M_u^{reflect}$;
8. 通过式(10)生成解题策略 π_t 和作答内容 a_t ;
9. 通过式(11)估计预测作答结果 \hat{r}_t' ;
10. 返回 π_t, a_t, \hat{r}_t' ;

4.1 教师信念

模拟教师在长期教学实践中形成的对学习者的学习表现与习题属性的稳定认知偏好可以通过持续观察学习者的答题过程, 逐步建立起对个体能力发展趋势和题目特征的综合判断机制。该认知机制在学习状态感知与个性化教学策略制定中起到关

键作用。因此, 模块从学习者历史作答数据中提取两类核心特征: 一方面是学习者的认知能力模式, 另一方面是习题的固有难度属性, 以支持教师智能体的动态决策与个性化配置。

为了刻画教师在判断学习者答题表现时的潜在认知机制, 借鉴“教师信念”(Teacher Belief)^[58]的建模策略, 不仅关注学习者能力与题目难度之间的匹配关系, 还融入教师对整体教学情境的经验认知, 包括学习者群体平均能力水平及测试整体难度的评估。

在建模前, 我们首先估计每位学习者的能力值 $\hat{\theta}_i$ 与每道题目的难度值 \hat{d}_e 。这些参数可借助外部工具, 如项目反应理论(Item Response Theory, IRT)、神经知识追踪模型或其他统计推断方法获得。这一设计旨在模拟真实教师在长期教学实践中形成的、对学习者的稳定认知偏好。引入 IRT 等外部工具并非要取代大语言模型的推理能力, 而是为其提供一个客观、稳健的初始认知锚点, 确保智能体后续的动态调整与反思机制并非无约束的零样本生成, 而是基于一个更贴近专家经验的信念基础展开。例如, 在 IRT 的二参数逻辑模型(2PL)中, 学习者 t 对题目 e 作答正确的概率定义为:

$$P_{te} = \frac{1}{1 + \exp(-a_e(\theta_t - b_e))} \quad (1)$$

其中, a_e 表示题目 e 的区分度, b_e 表示题目难度, θ_t 表示学习者能力。

此外, 为模拟教师对整体教学环境的经验性认知, 还计算学习者群体的平均能力及题目集合的考试难度指数:

$$\bar{\theta} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_i, \quad \bar{d} = \frac{1}{m} \sum_{e=1}^m \hat{d}_e \quad (2)$$

教师信念建模的核心是推断学习者是否具备完成特定题目的能力。该模块模拟教师基于学习者能力、题目难度及教学上下文, 对学习者的知识状态进行主观判断。具体实现中, 采用大语言模型(LLM)作为教师代理, 反映教师对学习者的答题成功的主观信念。

$$B_{t,e} = \text{LLM}^{\text{Tb}}(\hat{\theta}_i, \hat{d}_e, \bar{\theta}, \bar{d}, \hat{c}) \quad (3)$$

其中, $\text{LLM}^{\text{Tb}}(\cdot)$ 指在推理过程中融合了教师信念导向性提示的大语言模型, 从而具备识别与抽取教师信念的能力, \hat{c} 是教学上下文。

4.2 教师记忆模块

为实现对学习者的认知状态的动态建模与智能

响应,我们在 ACT-R^[59]认知架构的指导下进行了设计。结构化的教师记忆模块 M_u , 作为教师智能体持续感知学习者学习行为、组织反思性知识表征并驱动教学行为生成的核心组件。该模块通过周期性地提炼与优化记忆内容,过滤冗余观察、聚焦核心认知片段,从而有效管理长期行为信息,保证记忆表征的精确性与时效性。为此,该模块集成 4 个子模块:长期观察(Long-term Observation)、短期观察(Short-term Observation)、纠错反思(Corrective Reflection)与总结反思(Summary Reflection),并以大语言模型(LLM)作为认知映射器,实现行为轨迹与认知状态之间的转换。

4.2.1 长期观察

长期观察模块旨在帮助教师构建学习者稳定学习模式的全局认知背景,作为后续决策与认知推理的事实基础。设学习者在时间步 t 之前的全部学习行为序列为

$$H_u^{(1:t-1)} = \{(e_i, r_i)\}_{i=1}^{t-1} \quad (4)$$

其中, e_i 表示第 i 个学习任务的题目内容, $r_i \in \{0,1\}$ 表示学习者的作答结果(1 表示正确, 0 表示错误)。

4.2.2 短期观察

短期观察模块关注学习者在近期交互中的行为波动,捕捉其即时变化趋势,是支持教师进行个性化响应的重要输入。设定一个固定窗口大小 w , 系统构建最近 w 次交互序列如下:

$$H_u^{(t-w:t-1)} = \{(e_j, r_j)\}_{j=t-w}^{t-1} \quad (5)$$

其中, 本文遵循先前研究^[41]的经验, 将短期观察窗口大小设定为 $w = 5$, 以在即时性与行为稳定性之间取得平衡。

4.2.3 纠错反思机制

尽管大语言模型(LLM)具备先进的推理能力,但在判断或决策过程中仍可能出现偏差。例如, LLM 教师代理可能会给出一种解决方案,虽然逻辑上合理,但却超出了学习者的当前能力水平,或不符合其认知习惯与学习风格。为应对这类问题,我们在智能体中引入具备纠错功能的自我反思机制^[60],使其能够在预测学习者行为时主动识别自身判断与实际结果之间的差异^[61]。这种差异反映了智能体的元认知监控能力,使其能够对自身策略进行动态调整,从而实现更精准、个性化的支持。沿着这条路线,教师智能体配备了反思机制,用于监测其输出是否和学习者的能力水平和个性化特征相契合。

一旦识别出偏离或不适配之处,该机制会主动向决策模块发出调整信号,引导智能体重新审视并优化其策略选择。通过这种持续循环的自我校正过程,智能体在模拟学生思维路径时能够更加精确,有效降低因推理误差带来的干扰。为建模此类认知偏差并提供反馈支持,我们引入纠错反思机制。设教师对学习者当前任务 e_t 的预测结果为 \hat{r}_t^t , 实际作答结果为 r_t , 则定义预测偏差为

$$\delta_t = \hat{r}_t^t - r_t \quad (6)$$

系统以偏差值 δ_t 、当前任务内容 e_t 、教师信念状态 $B_{t,e}$ 以及短期观察 $H_u^{(t-w:t-1)}$ 作为输入生成纠错性认知反馈:

$$CR_t = \text{LLM}^{\text{Cr}}(\delta_t, e_t, B_{t,e}, H_u^{(t-w:t-1)}) \quad (7)$$

其中, CR_t 表示教师基于预测偏差所生成的个性化认知反馈。

4.2.4 总结反思机制

总结反思机制^[62]用于支持教师对学习者的认知状态进行高阶反思^[41]。为了帮助教师智能体在教学过程中形成对学习者的认知状态的深层理解,系统引入了一种总结性反思机制。该机制输入包括纠错反思结果 CR_t 、教师信念状态 $B_{t,q}$ 以及当前任务 e_t , 对学习者的表现进行阶段性归纳与概括。反思输出不仅揭示了学生当前的知识掌握情况,还揭示了其认知偏好与潜在障碍,正式输出可以表示为

$$SR_t = \text{LLM}^{\text{Sr}}(CR_t, B_{t,e}, e_t) \quad (8)$$

其中, SR_t 表示教师对学习者在时间步 t 所表现出的学习状态、认知倾向与潜在困难的总结性反思。

为实现在学习者认知状态演化的长期追踪与知识概念对齐建模,我们将阶段性总结结果 SR_t 及其对应的知识概念标签 kc_t 存入教师的长期反思记忆池 M_u^{reflect} 中,更新操作定义为

$$M_u^{\text{reflect}} \leftarrow M_u^{\text{reflect}} \cup \{(SR_t, kc_t)\} \quad (9)$$

其中, M_u^{reflect} 表示教师的反思池,用于记录各阶段认知总结及其关联知识点,为后续教学推理与策略生成提供支持。为避免其内容膨胀导致语言模型输入超限,系统引入摘要压缩机制,保留具有代表性的认知片段,以保障生成质量与计算效率。

4.3 教师的行动模块

为赋予教师智能体以类人化的教学引导与行为预测能力,2T-Agent 中为每个智能体设计了专门的行动模块。该模块在学习过程中能够基于当前观察与任务语境,动态调整行为策略,旨在实现个性化的学习支持。我们将教师智能体的行为建模为一

个“解释-认知”机制,以模拟真实教师在面对新任务时,如何结合认知记忆与教学经验,对学生可能的行为路径进行合理推断与干预。该模块聚焦于两个核心任务。

4.3.1 解释

给定学生的历史行为序列 $H_u^{(1:t-1)}$ 、教师反思池 M_u^{reflect} ,以及当前任务 e_t ,行动模块需推理学生在此任务中可能采取的解题策略 π_t ,并基于该策略生成相应的答题内容 a_t ,以形式化表述如下:

$$\pi_t, a_t = \text{LLM}^{\text{Int}}(M_u^{\text{reflect}}, e_t, H_u^{(1:t-1)}) \quad (10)$$

其中, π_t 表示语言化的解题思路, a_t 表示基于该思路推导出的模拟答案。该过程旨在模拟教师对学生可能思路的演绎与重构过程,作为对学生行为的先验解释基础。

4.3.2 认知预测

在获得解题思路与模拟作答后,行动模块进一步评估学生是否具备完成当前任务的能力,即模拟其对自身答题准确性的主观信心判断。该预测过程不仅关注结果的正确性,更强调教师对学生在认知层面的评估能力,具体建模如下:

$$\hat{r}_t' = \text{LLM}^{\text{Cog}}(e_t, \pi_t, a_t) \quad (11)$$

其中, \hat{r}_t' 表示教师对学生答题的预测结果,反映其在当前任务下的能力匹配程度。该机制体现了从“思路生成”到“信心判断”的完整认知路径,是教师智能体实现类人化教学推理的关键组成部分。

5 实验

本节聚焦于验证所提出框架的有效性。我们首先介绍实验中使用的真实数据集和对比方法,随后将分析研究的核心对象,即 2T-Agent。在与主流基线模型的系统性比较中,我们不仅关注其整体性能表现,更强调其在可解析性、冷启动方面和知识状态感知的独特优势。围绕这一目标,我们提出若干关键研究问题,并据此开展深入探讨,力求全面展现 2T-Agent 在认知建模方面的潜力与突破。

研究问题1:相较于现有方法,2T-Agent 在学习者表现预测方面展现出哪些性能改进与优势?

研究问题2:2T-Agent 的类教师思维生成的可解释结果是否符合教师认知逻辑?

研究问题3:2T-Agent 在开展认知状态评估时对外部诊断工具(如 CDM 等)的依赖程度如何?

研究问题4:在冷启动场景下,2T-Agent 是否能够有效完成表现预测?

研究问题5:2T-Agent 框架中各模块在整体性能中的作用如何?是否存在关键子模块对结果起主导作用?

研究问题6:2T-Agent 的认知层次感知能力如何?

研究问题7:相较于传统模型,2T-Agent 在诊断结果的可解释性方面具备哪些提升?

5.1 数据集

在 2T-Agent 框架下,为全面验证所提出模型的有效性,我们选了一个在学生认知建模方面具有丰富信息的基准数据集。该数据集由公开数据源与专家注释共同构建^[63],具有良好的代表性与研究价值。鉴于课程类型的多样性,我们将分析范围限定在机械力学(Mech),人工智能(AI),编程(Prog)三个子数据集上,以实现更集中的建模评估和更高的计算效率。表2展示了这三个子数据集的基本统计信息。

表2 数据集统计

统计/数据集	Mech	AI	Prog
练习记录数	22 286	1634	13 386
学习者数	754	126	294
练习题目数	137	22	120
知识概念数	162	27	131
平均练习长度	29.56	12.97	45.53
平均练习长度	29.56	12.97	45.53

5.2 实验设置

我们实验中的教师智能体(2T-Agent)基于大语言模型构建,核心模型包括本地部署的 DeepSeek-R1-7B,以及通过 OpenAI API 服务调用的 GPT-4o-mini 与 GPT-4o。为确保实验结果的确定性与可复现性,所有 GPT 模型的温度参数均设置为 0。除非特别说明,本文中 2T-Agent 默认使用 GPT-4o-mini;而在使用 GPT-4o 的验证性实验中,考虑到 API 成本,我们仅抽取了 100 名学习者进行模拟。

在训练与框架设置上,所有模型的网络参数采用 Xavier 方法进行初始化。我们遵循标准的机器学习流程,将数据集按照 70%(训练)、10%(验证)和 20%(测试)的比例进行随机划分。此外,我们框架中的短期观察窗口大小(w)固定为 5。所有实验的实现与评估均在统一的软硬件环境下进行,该环境为一台搭载 Intel Xeon Gold 5218 处理器(2.30 GHz)和四块 NVIDIA A100(40 GB)显卡的高性能服务器,操作系统为 64 位的 Ubuntu 20.04.5 LTS,计算

① <https://github.com/bigdata-ustc/EduCDM>

框架统一使用 PyTorch。

5.3 基准模型

为全面评估所提出智能体在判断学习者知识状态方面的有效性，我们选取了多个经典的认知诊断模型作为对比基准。所有对比模型均基于开源代码^①，从而保证比较结果的公平性与可信度。我们采用准确率(ACC)与 F1 分数(F1-Score)作为主要评估指标，以全面衡量各模型的性能表现。

(1)DKT^[9]: 该模型首次将深度学习技术引入知识追踪领域，采用长短期记忆网络(LSTM)对学生在不同知识领域的理解与掌握情况进行建模与预测，具有里程碑意义。

(2)GRKT^[64]: 该模型是一种基于图的知识追踪方法，它通过图神经网络探查知识点间的相互影响，并结合符合心理学的三阶段模型(知识检索、强化、学习/遗忘)，从而在更合理追踪学生知识演变的同时，提升了预测精度。

(3)DyGKT^[65]: 该模型通过构建动态图，捕捉了答题记录、时间间隔和实体关系这三大动态特性，革新了传统的知识追踪方法。

(4)THKT^[25]: 该模型首次将认知层次结构引入知识追踪任务，通过识别不同学生的目标认知层次并对各层次认知特征进行建模，实现了更具解释性的知识追踪结果。

(5)IRT^[8]: 项目反应理论模型通过分析学生不同题目上的作答表现，估计其潜在能力水平，是经典的测评理论之一。

(6)MIRT^[19]: 作为 IRT 的扩展模型，MIRT 能够同时评估学生在多个能力维度上的表现，适用于

多维认知结构的建模场景。

(7)DINA^[20]: 该模型假设每道题涉及多个知识点，并通过学生的答题情况推断其在各知识点上的掌握状态，采用二元分类方式对其进行描述，是认知诊断模型中的代表性方法之一。

(8)NCDM^[10]: 该模型引入神经网络结构，能够建模学生与试题之间更为复杂的交互关系，在提升认知诊断精度的同时，增强了模型的可解释性，是神经认知诊断的开创性工作。

(9)HCD^[66]: 该模型通过引入层次结构，对学生知识状态的变化进行有效约束，使其更贴合真实的教育场景。在当前数据集中，该模型与专家标注的知识层级结构较好匹配，有助于提升模型的学习与泛化能力。

(10)EduAgent^[44]: 该模型是一个生成式智能体框架，它将认知科学的先验知识融入大语言模型，引导其先进行推理再进行模拟，从而实现对动态学习行为的精准仿真。

5.4 性能预测(研究问题 1)

表 3 展示了我们提出的 2T-Agent 框架与多种主流基线模型在 Mech、AI 和 Prog 三个真实教育数据集上的性能对比，评估指标包括准确率(ACC)与 F1 分数(F1-Score)。需要注意的是，尽管我们并非以提升预测精度为最终目标，但通过在多数据集上的预测任务进行评估，可从多个维度间接验证 2T-Agent 在建模学习者认知状态方面的有效性。这种量化评估方法为进一步推进智能体在认知诊断、自适应教学与行为解释等方面的研究提供了坚实基础。

表 3 在 Mech、AI 和 Prog 数据集上不同方法的性能对比

Setting	Method	Mech-ACC	Mech-F1-Sco	AI-ACC	AI-F1-Score	Prog-ACC	Prog-F1-Scor
KT	DKT	79.63 ± 0.28	78.42 ± 0.30	77.13 ± 0.29	75.20 ± 0.27	72.80 ± 0.30	71.25 ± 0.28
	GRKT	80.03 ± 0.25	77.96 ± 0.28	78.62 ± 0.26	76.59 ± 0.29	77.44 ± 0.23	76.59 ± 0.25
	DyGKT	81.35 ± 0.19	79.13 ± 0.21	79.22 ± 0.24	78.49 ± 0.22	78.47 ± 0.18	77.03 ± 0.20
	THKT	<u>82.16 ± 0.18</u>	<u>79.95 ± 0.20</u>	78.54 ± 0.25	<u>79.46 ± 0.19</u>	74.11 ± 0.27	73.43 ± 0.29
CD	IRT	74.56 ± 0.30	73.28 ± 0.28	77.29 ± 0.27	73.36 ± 0.30	73.42 ± 0.28	72.17 ± 0.29
	MIRT	76.41 ± 0.26	75.32 ± 0.29	78.99 ± 0.24	75.03 ± 0.28	74.13 ± 0.25	72.73 ± 0.27
	DINA	74.26 ± 0.29	71.35 ± 0.30	78.21 ± 0.28	74.07 ± 0.29	76.15 ± 0.24	74.88 ± 0.26
	NCDM	80.26 ± 0.21	78.99 ± 0.23	<u>81.88 ± 0.17</u>	79.03 ± 0.19	76.45 ± 0.22	75.25 ± 0.24
	HCD	81.58 ± 0.19	79.83 ± 0.21	80.85 ± 0.20	78.52 ± 0.21	<u>79.61 ± 0.17</u>	<u>77.73 ± 0.19</u>
Agent	EduAgent (GPT-4o-mini)	80.26 ± 0.12	76.47 ± 0.14	79.32 ± 0.20	77.92 ± 0.05	78.46 ± 0.18	74.93 ± 0.16
	EduAgent (GPT-4o(100))	79.45 ± 0.14	77.21 ± 0.06	79.11 ± 0.15	77.07 ± 0.13	78.51 ± 0.17	74.28 ± 0.08
Ours	2T-Agent (DeepSeek-R1)	79.56 ± 0.16	78.44 ± 0.18	80.49 ± 0.14	78.60 ± 0.17	77.46 ± 0.19	76.93 ± 0.20
	2T-Agent (GPT-4o-mini)	83.12* ± 0.08	80.73* ± 0.09	82.11 ± 0.10	80.29 ± 0.11	83.26* ± 0.04	84.93* ± 0.02
	2T-Agent (GPT-4o(100))	82.26 ± 0.09	80.49 ± 0.10	83.89* ± 0.03	81.31* ± 0.05	82.26 ± 0.06	± 0.05

注: 所有数值均为 5 折交叉验证的均值 ± 95% 置信区间半宽。最佳结果以加粗表示，次优结果以下划线标示。*表示加粗结果相较于次优结果的提升，在双尾配对 *t* 检验中具有统计显著性 ($p < 0.05$)。

从对比结果来看, 2T-Agent 框架在各项任务上均展现出显著优势, 整体优于传统的知识追踪(KT)方法与认知诊断(CD)模型。此外, 与同为教育智能体模型的 EduAgent 相比, 当采用相同的大语言模型(如 GPT-4o-mini)作为基座时, 我们的 2T-Agent 模型在各项评估指标上均展现出更优越的性能。其中, 2T-Agent(GPT-4o-mini)在六个指标中有四项取得最佳结果, 特别是在 Prog 数据集上达到 83.26% 的准确率和 84.93% 的 F1-Score, 显著超过其他模型($p < 0.05$)。此外, 2T-Agent(GPT-4o(100))仅在 GPT-4o 驱动下对随机选取的 100 名学生进行了建模分析, 旨在控制成本的前提下进行效果验证。尽管数据量相对较小, 其在 AI 数据集上依然取得了最高的 AI-ACC(83.89%)与 AI-F1-Score(81.31%)指标。这一结果不仅印证了模型在小样本场景下的稳健性, 也进一步强调了高性能大语言模型在教师智能体中的关键作用。进一步观察发现, 驱动教师智能体的大语言模型能力对整体表现具有关键影响。不同规模与能力的 LLM 在反思生成与认知建模过程中展现出不同水平的支持能力, 凸显了大模型在理解、归因和决策生成等方面的潜力。这也提示我们, 构建具备类人认知能力的教学智能体, 不仅需要合理的系统结构设计, 还应充分发挥语言模型在多层次认知推理中的优势。最后还发现, 从 95% 置信区间半宽可见, 大模型类方法(特别是 2T-Agent)在三个数据集上的性能波动显著低于传统 KT/CD 模型, 这表明引入大模型推理能力后, 预测结果在不同测试集划分下更加一致, 模型的稳健性与泛化能力得到明显提升。

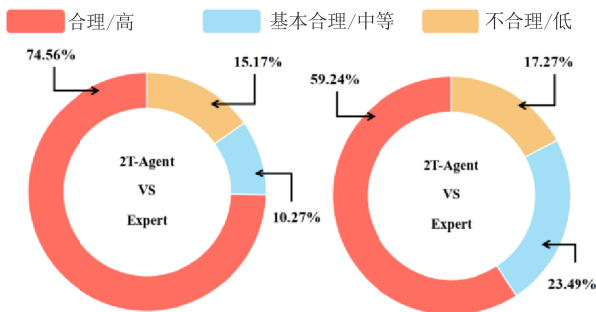
5.5 2T-Agent 生成的解释的质量评估(研究问题 2)

为深入验证本文提出的 2T-Agent 在模拟教师思维模式过程中的合理性与专业性, 我们设计并执行了一项严格的专家评估实验。

5.5.1 评估流程与量规

我们邀请了三位在计算机教育与认知诊断领域具备 5 年以上研究或教学经验的领域专家参与本次评估。评估材料为采用本文提出的 2T-Agent 模型, 针对 Prog 数据集中随机抽取的 20 名学习者的共 200 条答题记录所生成的解释文本。为确保评估的客观性与公平性, 我们遵循了严格的独立盲审流程。首先, 我们向三位专家提供了详细的评分量规(见附录表 5)并进行了培训, 以确保图 3 专家对他

们对各项评分标准有共同且一致的理解。随后, 我们将这 200 条由 2T-Agent 生成的解释文本与对应的题目、学生答题情况一同呈现给专家。三位专家在互不知晓对方评分的情况下, 进行了独立的背对背评分。



(a) 专家对 2T-Agent 生成解释合理性的评价 (b) 专家评价 2T-Agent 生成解释的专业性

图 3 专家对 2T-Agent 生成解释的合理性与专业性评估

5.5.2 一致性检验与结果分析

在收集到所有评分后, 我们首先检验了三位专家评分结果的一致性。我们采用了适用于多位评分者的弗莱斯 Kappa 系数来衡量评分者信度。计算得出的 Kappa 值为 0.796, 根据 Landis and Koch^[67]的经典标准, 该值表明专家们的评分具有实质性一致性, 从而证实了本次评估结果的可靠性与可信度。

在通过一致性检验确认了评估的可靠性后, 我们对所有评分进行了汇总, 结果如图 3 所示。统计结果显示, 2T-Agent 所生成解释的合理性获得了专家的高度认可, 被评为“高”或“中等”合理性的比例合计达到了 84.83%, 这表明该智能体在多数情况下能够生成与人类专家逻辑相近、过程清晰的解题思路。在专业性维度上, 59.24% 的生成结果被认为具备较高的专业性, 体现出 2T-Agent 在捕捉关键知识点、组织解题语言和呈现教师意图方面具备了初步的教学理解能力。与此同时, 17.27% 的结果被评为专业性不足, 也提示模型在术语使用、深度引导或学科表达方面仍有改进空间。该评估结果直接验证了“解释与认知链生成模块”设计的有效性, 构成了系统机制与实验结果间的相互印证。这种生成高质量、类教师解释的能力, 是增强模型可解释性、提升用户信任并构建智能体—学习者互动闭环的关键支撑。

5.6 2T-Agent 的工具依赖性分析(研究问题 3)

为了探究不同认知诊断工具对本文提出的 2T-Agent 在知识状态预测能力上的影响, 我们随机选取了两位学习者, 并在 Mech 数据集上比较了基

于历史统计(HStat)、IRT(Item Response Theory)与DKT(Deep Knowledge Tracing)三种诊断工具的

下的2T-Agent的应用对比结果,效果如图4所示,对比结果呈现出若干值得关注的现象。

S ₁	2T-Agent (HStat)	×	✓	✓	✓	✓	×	✓	✓	×	×	✓	×	✓	×	✓
	2T-Agent (IRT)	✓	×	×	✓	×	✓	✓	✓	✓	✓	✓	×	✓	×	✓
	2T-Agent (DKT)	×	×	✓	✓	✓	✓	✓	×	×	×	✓	×	✓	×	✓
S ₂	Actual Responses	0	1	1	1	0	1	1	1	0	0	1	1	0	1	1
	Exercises	19	20	21	22	23	27	28	32	44	45	62	64	73	77	81
	Actual Responses	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
S ₂	2T-Agent (HStat)	✓	×	✓	✓	×	×	✓	✓	×	✓	✓	×	✓	✓	✓
	2T-Agent (IRT)	✓	×	✓	✓	✓	×	×	✓	×	✓	✓	×	✓	✓	✓
	2T-Agent (DKT)	×	✓	×	✓	✓	×	✓	✓	×	×	✓	×	✓	✓	×

图4 2T-Agent 基于不同知识状态诊断工具的预测结果(✓表示智能体预测学生会答对, ×表示预测智能体预测学生会错误), HStat代表基于历史的统计

首先,在学习过程的后半阶段,2T-Agent在三种诊断工具辅助下的预测结果逐渐趋于一致,表现出较强的稳定性与鲁棒性。这表明随着练习的持续深入,诊断工具对预测结果的影响逐渐减弱,教师智能体在评估学生知识状态时愈发依赖自身的学习与判断,而非外部工具的指导,从侧面体现出教师智能体“智能性”的逐步增强。其次,在学习初始阶段的预测结果中可以看出,2T-Agent对诊断工具存在较强依赖,工具的输出在很大程度上决定了智能体的判断。例如,学习者S₁和S₂在前两个题目(第19题和第20题)上的预测结果几乎完全受诊断工具的引导,不同工具产生了明显不同的预测输出。这一趋势与2T-Agent的设计机制一致,即通过持续学习和反馈机制不断优化其自主预测能力。第三,我们也观察到2T-Agent在某些情境下的局限性。例如,学习者S₁在题目64上,以及学习者S₂在题目32上的预测结果,在三种诊断工具辅助下均与实际作答情况相反。这表明模型在某些情况下仍无法全面捕捉学生行为的复杂性,预测偏差可能并非源于知识掌握程度本身,而是受到心理状态、操作失误等尚未建模因素的影响。最后,尽管2T-Agent在早期预测中较为依赖诊断工具,不同学生的教师智能体仍能够根据个体特征进行有效区分,保证预测结果是基于该学生自身的表现做出。例如,在题目23上,2T-Agent对学习者的预测展现出明显的个体差异性。尽管学生的真实作答情况仍可能受到未建模潜在因素的干扰,但这一从“外部依赖”到“自

主判断”的转变,仍然直接验证了“教师记忆与反思模块”的有效性,证明其通过持续学习与自我修正成功构建了稳健的内部认知模型。具体而言,该自适应过程由“纠偏反思”与“总结反思”机制协同驱动:前者负责修正即时预测偏差,后者则将修正经验归纳并更新至长期记忆,从而实现对初始偏差的自主修正。

5.7 冷启动表现预测能力(研究问题4)

冷启动实验旨在回答一个关键问题:在缺乏学生历史交互数据的情境下,2T-Agent是否能够有效完成对学生表现的认知建模与预测?为此,我们设计了模拟冷启动环境的实验场景,以系统评估不同模型在零交互条件下的泛化能力和推理能力。为了符合冷启动设定,我们在Mech与Prog两个数据集上对学生数据进行处理,仅保留每位学习者在该领域首次出现时的一小部分交互记录,供模型进行初始化,其余交互记录用于性能评估。在这种极端数据稀缺的条件下,模型无法依赖完整的学生行为序列,而必须通过有限信息完成认知状态建模与表现预测,这对模型的推理能力和泛化能力提出了更高要求。

如表4所示,有几个有意思的实验结果。首先,2T-Agent在Mech与Prog两个数据集上的表现均显著优于传统模型,验证了其在零交互条件下对学生认知状态的建模与表现预测能力。其次,可以观察到知识追踪模型DKT的整体表现优于IRT和NCDM,这可能得益于其基本建模范式更强调学生

行为的时间序列特征, 从而具备更强的泛化能力。相比之下, 认知诊断模型如 IRT 和 NCDM 通常依赖对每位学生的充分训练才能展现优势, 因此在冷启动情境下适应性较弱。最后, 进一步比较表 3 中 2T-Agent 在两个数据集上的表现可以发现, 其性能并未出现如其他模型那样的断崖式下降, 反而保持高度一致, 说明通过大语言模型模拟教师思维, 能够有效增强模型在冷启动环境中的稳健性和推理能力。

表 4 Mech 和 Prog 数据集上 F1-Score 和 ACC 的冷启动对比分析

Model	Mech		Prog	
	ACC	F1-Score	ACC	F1-Score
DKT	68.78	66.80	59.75	58.78
IRT	55.72	53.45	54.27	52.03
NCDM	61.85	60.87	59.52	58.21
2T-Agent	78.81	77.33	77.41	78.92

5.8 消融实验(研究问题 5)

为评估 2T-Agent 各组成模块的功能贡献, 我们设计了四种消融设置, 分别移除教师信念(w/o Tb)、教师记忆(w/o Mem)、反思机制(w/o Ref)以及解释模块(w/o Int), 并在多领域数据集上进行对比实验。实验结果如图 5 所示。首先可以观察到完整模型在所有任务中均取得最优性能, 验证了多模块协同机制的有效性; 特别是在具有强序列依赖特性的领域, 完整模型优势更为显著, 表明跨模块信息流动对复杂认知任务的推理至关重要。

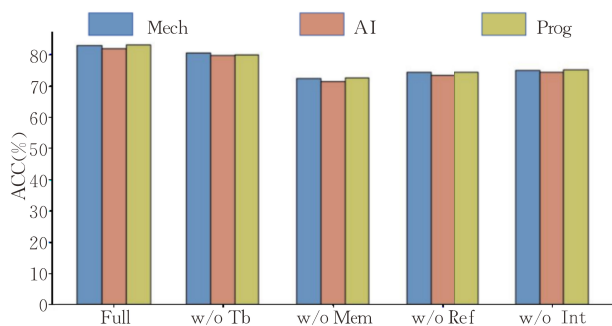


图 5 在三个数据集上的消融实验对比

其次, 模块贡献度分析显示: 教师信念模块(Tb)的消融虽造成一定性能损失, 但表明大语言模型凭借其预训练知识与内在推理能力, 可在缺乏明确领域先验的情况下部分维持模型效能。这一结果印证了 LLM 在教育场景中的知识泛化潜力, 但其局限性也凸显了领域特异性信念建模的必要性。教师记忆模块(Mem)在持续建模学生历史表现中发挥核

心作用, 其缺失导致模型对学生知识状态的动态推断能力显著下降。这体现了历史行为轨迹对当前推理决策的支撑作用, 尤其在知识累积型任务(如 Mech)中表现尤为明显。反思机制(Ref)对高阶认知任务的推理深度具有重要影响, 其缺失导致模型在分析、评价等需要元认知能力的任务中表现下滑。这凸显了偏差检测与修正机制对认知诊断精度的提升作用, 特别是在处理复杂问题解决过程中的认知迭代。解释模块(Int)则在关键行为特征识别与解释连贯性方面发挥关键作用, 其消融不仅削弱了模型对学生行为模式的捕捉能力, 也降低了诊断结果的可解释性。这反映了解释生成过程对特征选择与决策过程的优化效应, 尤其在需要人机协作的教育场景中具有重要价值。

最后, 模块贡献亦呈现领域敏感性: 在知识结构相对清晰的领域(如 Mech), 教师信念模块的贡献度较低, 而在需要深度领域知识融合的场景(如 AI)中其作用更为凸显; 记忆模块在具有强序列依赖特性的任务(如编程)中重要性显著提升, 体现了不同认知任务对历史信息利用程度的差异。这种功能耦合性印证了人类教师认知过程的整体性特征, 即观察、判断、反思与反馈等环节的相互依赖与协同。

5.9 2T-Agent 的认知层次感知能力分析(研究问题 6)

为验证所提出的 2T-Agent 在认知层面模拟教师思维的能力, 我们基于 Prog 编程教育数据集设计了布鲁姆(Bloom)认知分类预测实验。该数据集由编程学习者在真实环境中的作答记录构成, 每道题目均由教育专家依据布鲁姆分类法^[68]进行认知层次标注, 涵盖记忆、理解、应用、分析、评价与创造六个层级。在本实验中, 我们随机选取了 20 名学习者的作答数据作为测试集, 并让模型预测每道题目的布鲁姆认知层次标签, 以此衡量 2T-Agent 对题目认知要求的理解与判断能力。

图 6 展示了预测混淆矩阵结果: 模型在理解、应用、分析和评价四个层级上的预测准确率分别达到 83.5%、82.1%、84.6%和 82.4%, 表明 2T-Agent 能有效捕捉题目认知特征, 减轻专家标注工作量。

相邻层级存在一定混淆, 例如“理解”易被误判为“应用”(7.6%), “应用”易被误判为“分析”

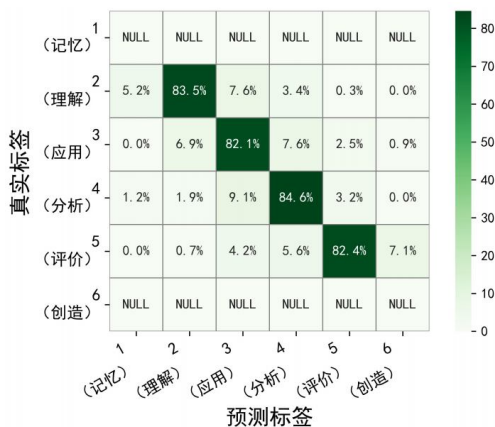


图6 2T-Agent 预测的布鲁姆认知层级的混淆矩阵(本实验所用原始数据集的专家标注仅覆盖“理解”至“评价”四个认知层次,故“记忆”与“创造”层次无对应数据)

(7.6%), 反映出处理边界模糊认知阶段的挑战, 同时符合 Bloom 层次递进式结构特性。远距离层级误判极少, 例如“理解”很少被预测为“评价”或“创造”, 说明模型对整体认知层级保持良好认知。

5.10 案例分析(研究问题 7)

我们从 AI 数据集中选取了一位学习者的作答记录, 分析其在一道涉及函数推理任务($F(X)=Y$ 形式) 的表现。该任务要求学生在已知输入 X 和输出 Y 的前提下, 推理出未知的函数 F , 以理解背后的规则关系。针对该任务, 图 7 对比了三种不同模型的预测结果与解释能力: 传统的神经认知诊断模型(NCDM), 以及基于大语言模型进行思维模拟的 EduAgent(学生视角)与 2T-Agent 框架(教师视角)。

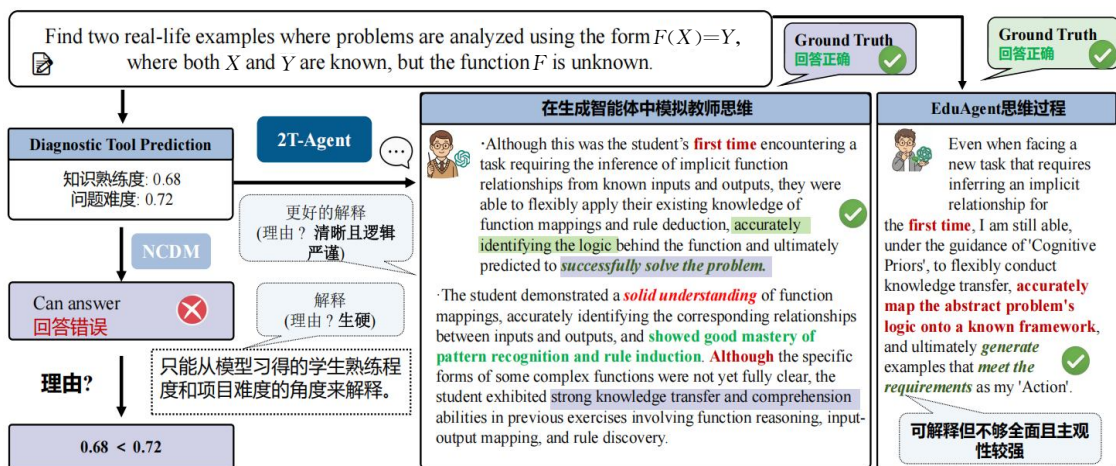


图7 2T-Agent 与 EduAgent 和 NCDM 的可解性案例分析

从图中可以观察到, NCDM 仅依赖学生的知识掌握度(0.68)与题目难度(0.72)之间的静态匹配, 预测该学生无法正确完成该函数推理任务。然而, 该学生实际成功解答, 说明基于传统可解性的模型无法充分刻画学生在推理类任务中所展现出的迁移思维与关系抽象能力。

而 EduAgent 则从学生的第一视角出发, 通过其“反思”(Reflection)机制模拟了解决该问题的完整认知路径: 它并非直接匹配知识点, 而是将抽象的 $F(X)=Y$ 问题成功映射到了其知识库中具体的“已知框架”上。这一过程直接生成了最终的正确答案, 从而具象化地展现了该学生是如何运用迁移和推理能力来解决一个全新任务的。然而, 这一思维过程虽然可解释, 但也存在不够全面且主观性较强的特点。

相比之下, 2T-Agent 通过模拟教师视角, 结合学生在人工智能基础课程中涉及的函数映射、模式

识别、规则推导与归纳推理等相关知识点的作答轨迹, 识别出其在结构识别与模式提取方面的稳定认知优势。尽管该学生此前未直接接触过 $F(X)=Y$ 类函数反推任务, 但其在相关题目中展现出的规则归纳与逻辑推理能力被 2T-Agent 有效捕捉并迁移应用, 从而做出了合理判断。该过程展现了 2T-Agent 对“解题可行性”的更深层建模: 不仅考虑掌握的知识点, 更融合其在上下文任务中的策略迁移、认知迁移和抽象推理能力, 从而提供更具解释性与预测力的诊断结果。该案例表明, 2T-Agent 能够有效模拟学生在人工智能抽象任务中的认知路径, 显著提升对推理型问题的适应能力与判断精度。

6 结论与局限性

本文提出了 2T-Agent 框架, 一个具备类教师认知机制的智能体认知诊断框架。该框架通过模拟教师在真实教学场景中的认知判断过程, 系统建构

了任务感知的信念建模、基于行为轨迹的观察反思机制,以及因果驱动的解释生成模块,从而实现对学生认知状态的动态建模与偏差诊断。实验结果表明,2T-Agent在多个认知评估任务中均取得领先性能,展现出以下显著优势:一是具备良好的冷启动能力,即使在缺乏完整历史行为数据的情况下也能做出合理判断;二是具备较强的可解释性,能够输出“知识状态—行为过程—任务表现”的因果链条,支持精准教学干预;三是具备对学生思维层级的判别能力,能够辅助完成布鲁姆认知分类,为任务分层与教学设计提供有效支撑。

尽管当前框架在认知评估中展现出显著优势,仍存在两方面技术瓶颈:其一,反思与解释模块对创造性推理等高阶认知行为的建模采用简化策略,尚未完全覆盖认知过程的非线性动态关系;其二,核心模块高度依赖大语言模型的推理能力,在低资源语言教育场景或专业领域中易因模型知识截止或术语理解偏差导致诊断误差。其三,当前框架未能有效区分由知识欠缺导致的认知性误差,与因操作失误、情绪波动或猜测等引发的非认知性误差。其四,框架在真实应用中需应对多重挑战。在性能层面,大语言模型的推理能力直接影响诊断的精度,其生成解释的忠实性也需验证,同时响应延迟可能影响实时交互。在工程层面,需要考虑大模型的计算成本,以及专属智能体设计带来的可扩展性问题,这些都对大规模部署的资源与架构设计提出了更高要求。最后,本研究的验证虽然基于源于真实教育场景的离线数据集,但仍缺乏在动态、实时的教学交互中的应用验证。针对上述局限,未来研究可聚焦多智能体协同框架的拓展,通过引入教师智能体、学生难题及策略协调机制,构建具备多角色互动能力的认知评估系统,从而在群体协同诊断、个性化教学反馈等场景中提升智能体的推理实效性与场景适应性。

参考文献

- [1] Liu Q, Huang Z, Yin Y, et al. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 33(1): 100-115
- [2] Sun X, Zhang K, Shen S, et al. Mhed: Multi-hierarchy interactive constraint-aware cognitive diagnosis framework. *Expert Systems with Applications*, 2025, 283: 127701
- [3] Sun X, Liu Q, Zhang K, et al. LGS-KT: Integrating logical and grammatical skills for effective programming knowledge tracing. *Neural Networks*, 2025, 185: 107164
- [4] Chen B, Zheng W, Zhao L, et al. Leveraging large language models to assist philosophical counseling: Prospective techniques, value, and challenges. *Humanities and Social Sciences Communications*, 2025, 12(1): 1-15
- [5] Sun X, Liu Q, Zhang K, et al. Harnessing code domain insights: Enhancing programming knowledge tracing with large language models. *Knowledge-Based Systems*, 2025, 317: 113396
- [6] Shen S, Liu Q, Huang Z, et al. A survey of knowledge tracing: Models, variants, and applications. *IEEE Transactions on Learning Technologies*, 2024, 17: 1858-1879
- [7] Wang F, Gao W, Liu Q, et al. A survey of models for cognitive diagnosis: New developments and future directions. *arXiv preprint arXiv:2407.05458*, 2024
- [8] Lord F M. *Applications of item response theory to practical testing problems*. New York, USA: Routledge, 2012.
- [9] Piech C, Bassen J, Huang J, et al. Deep knowledge tracing. *Advances in Neural Information Processing Systems*, 2015, 28: 1-9
- [10] Wang F, Liu Q, Chen E, et al. NeuralCD: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 2022, 35(8):8312-8327
- [11] Zhuang Y, Liu Q, Huang Z, et al. Fully adaptive framework: Neural computerized adaptive testing for online education//*Proceedings of the AAAI Conference on Artificial Intelligence*. Virtual, 2022: 4734-4742
- [12] Lv R, Liu Q, Gao W, et al. GenAL: Generative agent for adaptive learning//*Proceedings of the AAAI Conference on Artificial Intelligence*. Philadelphia, USA, 2025: 577-585
- [13] Sun X, Zhang K, Liu Q, et al. DASKT: A dynamic affect simulation method for knowledge tracing. *IEEE Transactions on Knowledge and Data Engineering*, 2025, 37(4): 1714-1727
- [14] Gao W, Liu Q, Huang Z, et al. RCD: Relation map driven cognitive diagnosis for intelligent education systems//*Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Virtual, 2021: 501-510
- [15] Gao W, Liu Q, Wang H, et al. Zero-1-to-3: Domain-level zero-shot cognitive diagnosis via one batch of early-bird students towards three diagnostic objectives//*Proceedings of the AAAI Conference on Artificial Intelligence*. Vancouver, Canada, 2024: 8417-8426
- [16] Ke Y H, Yang R, Lie S A, et al. Enhancing diagnostic accuracy through multi-agent conversations: using large language models to mitigate cognitive bias. *arXiv preprint arXiv:2401.14589*, 2024
- [17] Qi T, Ren M, Guo L, et al. ICD: A new interpretable cognitive diagnosis model for intelligent tutor systems. *Expert Systems with Applications*, 2023, 215: 119309
- [18] Chen X, Wu L, Liu F, et al. Disentangling cognitive diagnosis with limited exercise labels. *Advances in Neural Information Processing Systems*, 2023, 36: 18028-18045
- [19] Reckase M D. *Multidimensional Item Response Theory*. New York, USA: Springer, 2009: 79-112
- [20] De La Torre J. DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, 2009, 34(1): 115-130
- [21] Zhang M, Wang Z, Xing R, et al. Disentangling heterogeneous knowledge concept embedding for cognitive diagnosis on untested knowledge. *arXiv preprint arXiv:2405.16003*, 2024

- [22] Wang S, Zeng Z, Yang X, et al. Boosting neural cognitive diagnosis with student's affective state modeling//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 620-627
- [23] Zhang Z, Wu L, Liu Q, et al. Understanding and improving fairness in cognitive diagnosis. *Science China Information Sciences*, 2024, 67(5): 152106
- [24] Ma D, He F, Yue Y, et al. Graph convolutional networks for street network analysis with a case study of Urban Polycentricity in Chinese Cities. *International Journal of Geographical Information Science*, 2024, 38(5): 931-955
- [25] Sun X, Zhang K, Shen S, et al. Target hierarchy-guided knowledge tracing: Fine-grained knowledge state modeling. *Expert Systems with Applications*, 2024, 251: 123898
- [26] Zanellati A, Di Mitri D, Gabbriellini M, et al. Hybrid models for knowledge tracing: A systematic literature review. *IEEE Transactions on Learning Technologies*, 2024, 17: 1021-1036
- [27] Cui C, Yao Y, Zhang C, et al. DGEKT: A dual graph ensemble learning method for knowledge tracing. *ACM Transactions on Information Systems*, 2024, 42(3): 1-24
- [28] Wang F, Huang Z, Liu Q, et al. Dynamic cognitive diagnosis: An educational priors-enhanced deep knowledge tracing perspective. *IEEE Transactions on Learning Technologies*, 2023, 16(3): 306-323
- [29] Sun X, Liu Q, Zhang K, et al. Tracking knowledge state transitions in cognitive diagnosis for an exercise-exam learning model. *IEEE Transactions on Industrial Informatics*, 2025.21(10): 8012-8022
- [30] Zhu Tianyu, Huang Zhenya, Chen Enhong, et al. Personalized test recommendation method based on cognitive diagnosis. *Chinese Journal of Computers*, 2017, 40(1): 176-191(in Chinese)
(朱天宇, 黄振亚, 陈恩红等. 基于认知诊断的个性化试题推荐方法. *计算机学报*, 2017, 40(1): 176-191)
- [31] Jin Tiancheng, Dou Liang, Xiao Chunyun, et al. Personalized OJ exercise recommendation method integrating memory and cognition. *Chinese Journal of Computers*, 2023, 46(1): 103-124(in Chinese)
(金天成, 窦亮, 肖春芸等. 记忆与认知融合的个性化OJ习题推荐方法. *计算机学报*, 2023, 46(1): 103-124)
- [32] Bail C A. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 2024, 121(21): e2314021121
- [33] Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024, 18(6): 186345
- [34] Li Peng, Chen Shaofei, Yi Chushu, et al. A survey on multi-agent risk decision-making theories and methods. *Chinese Journal of Computers*, 2025: 1-33(in Chinese)
(李鹏, 陈少飞, 易楚舒等. 多智能体风险决策理论与方法研究综述. *计算机学报*, 2025: 1-33)
- [35] Sun X, Liu Q, Zhang K, et al. Plug-and-play fine grained neural cognitive diagnosis framework. *Engineering Applications of Artificial Intelligence*, 2025, 162: 112276
- [36] Zhang A, Chen Y, Sheng L, et al. On generative agents in recommendation//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, USA, 2024: 1807-1817
- [37] Sun Y Z, Sun H L, Ma J C, et al. Multimodal Agent AI: A survey of recent advances and future directions. *Journal of Computer Science and Technology*, 2025, 40(4): 1046-1063
- [38] Tran K T, Dao D, Nguyen M D, et al. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025
- [39] Wang Luqiao, Zhou Yangtao, Li Qingshan, et al. Reviewer recommendation for collaborative code review based on large language models. *Journal of Software*, 2025, 36(6): 2558-2575(in Chinese)
(王路桥, 周洋涛, 李青山等. 基于大语言模型的多智能体协作代码评审人推荐. *软件学报*, 2025, 36(6): 2558-2575)
- [40] Sun X, Zhang K, Liu Q, et al. Harnessing domain insights: A prompt knowledge tuning method for aspect-based sentiment analysis. *Knowledge-Based Systems*, 2024, 298: 111975
- [41] Gao W, Liu Q, Yue L, et al. Agent4Edu: Generating learner response data by generative agents for intelligent education systems//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(22): 23923-23932
- [42] Liu N, Wang Z, Baraniuk R, et al. Open-ended knowledge tracing for computer science education//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 3849-3862
- [43] Liu J, Huang Z, Xiao T, et al. SocraticLM: Exploring socratic personalized teaching with large language models. *Advances in Neural Information Processing Systems*, 2024, 37: 85693-85721
- [44] Xu S, Zhang X, Qin L. Eduagent: Generative student agents in learning. *arXiv preprint arXiv:2404.07963*, 2024
- [45] Valmeekam K, Marquez M, Olmo A, et al. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 2023, 36: 38975-38987
- [46] Hadi M U, Qureshi R, Shah A, et al. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2023, 1(3): 1-26
- [47] Yao S, Yu D, Zhao J, et al. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 2023, 36: 11809-11822
- [48] Yan Y, Wwen H, Zhong S, et al. UrbanCLIP: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web//Proceedings of the ACM Web Conference 2024. Singapore, 2024: 4006-4017
- [49] Liang X, Wang B, Huang H, et al. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *arXiv preprint arXiv:2304.13343*, 2023
- [50] Qin Y, Liang S, Ye Y, et al. ToolLLM: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023
- [51] Park J S, O' brien J, Cai C J, et al. Generative Agents: Interactive simulacra of human behavior//Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology. San Francisco, USA, 2023: 1-22
- [52] Dan Y, Lei Z, Gu Y, et al. EduChat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*, 2023

- [53] Kieser F, Wulff P, Kuhn J, et al. Educational data augmentation in physics education research using ChatGPT. *Physical Review Physics Education Research*, 2023, 19(2): 020150
- [54] Durante Z, Huang Q, Wake N, et al. Agent AI: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024
- [55] Xie J, Zhang K, Chen J, et al. Revealing the barriers of language agents in planning. *arXiv preprint arXiv:2410.12409*, 2024
- [56] Sikula J. *Handbook of Research on Teacher Education*. New York, USA: Simon & Schuster Macmillan, 1996
- [57] Glaser R, Chudowsky N, Pellegrino J W. *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, USA: National Academies Press, 2001
- [58] Fang Z. A review of research on teacher beliefs and practices. *Educational Research*, 1996, 38(1): 47-65
- [59] Anderson J R, Lebiere C J. *The atomic components of thought*. Hove, UK: Psychology Press, 1998
- [60] Ellis R. A typology of written corrective feedback types. *ELT Journal*, 2009, 63(2): 97-107
- [61] Zhan Y, Liu Q, Gao W, et al. CoderAgent: Simulating student behavior for personalized programming learning with large language models. *arXiv preprint arXiv:2505.20642*, 2025
- [62] Baron J. Reflective thinking as a goal of education. *Intelligence*, 1981, 5(4): 291-309
- [63] Yu J, Wang Y, Zhong Q, et al. MOOCCubeX: A large knowledge-centered repository for adaptive learning in moocs//*Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. Virtual, 2021: 4643-4652
- [64] Cui J, Qian H, Jiang B, et al. Leveraging pedagogical theories to understand student learning process with graph-based reasonable knowledge tracing//*Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain, 2024: 502-513
- [65] Cheng K, Peng L, Wang P, et al. DyGKT: Dynamic graph learning for knowledge tracing//*Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona, Spain, 2024: 409-420
- [66] Sun X, Liu Q, Zhang K, et al. HCD: A hierarchy constraint-aware neural cognitive diagnosis framework. *Neural Networks*, 2025, 190: 107668
- [67] Landis J R, Koch G G. The measurement of observer agreement for categorical data. *Biometrics*, 1977, 33: 159-174
- [68] Travers R M W. Taxonomies of educational objectives and theories of classification. *Educational Evaluation and Policy Analysis*, 1980, 2(2): 5-23

附录 1.

表5 专家评估的评分量规

评估维度	分数等级	评分标准	评估维度	分数等级	评分标准
合理性 (Reasonableness)	高(High/3分)	解释完全符合逻辑, 清晰易懂, 且高度契合学生的潜在思维路径。	专业性 (Professionalism)	高(High/3)	解释准确使用教育学和学术术语, 体现深刻的教学洞察力, 并能提供有价值的干预建议。
	中(Medium/2分)	解释基本符合逻辑, 但存在少量歧义, 或与学生思维的关联性稍弱。		中(Medium/2分)	解释在专业性上基本正确, 但缺乏深度的教学洞察, 或术语使用不够精确。
	低(Low/1分)	解释不符合逻辑、难以理解, 或与学生问题完全无关。		低(Low/1分)	解释使用了错误的术语, 或完全不具备教学上的专业价值。



SUN Xin-Jie, Ph.D. candidate. His research interests include intelligent education and adaptive learning.

LIU Qi, Ph.D., professor. His research interests focus on data mining.

Background

This study focuses on the challenge of student cognition assessment in intelligent educational systems, which represents a core research direction in the field of Education Data Mining and Cognitive Diagnosis. Current mainstream international approaches predominantly rely on modeling question-

ZHANG Kai, Ph.D., associate professor. His research interests focus on large model optimization.

GONG Wei-Yin, Ph.D. candidate. His research interests focus on multimodal intent recognition.

SHEN Shuang-Hong, Ph.D., associate professor. His research interests focus on intelligent education.

WANG Fei, Ph.D., postdoctoral fellow. His research interests focus on cognitive diagnosis.

answering performance, yet they universally overlook the influences of cognitive motivations, behavioral evolution patterns, and contextual factors in the learning process. Although deep knowledge tracing models (e.g., DKT) and neural cognitive diagnosis frameworks (e.g., NCDM) have

achieved phased progress, they still exhibit three core limitations in complex teaching scenarios: insufficient adaptability, limited depth in cognitive characterization, and weak interpretability, making it difficult to construct a causal explanation chain of “knowledge state-behavior process-task performance.”

The 2T-Agent framework proposed in this paper achieves key breakthroughs by emulating human teachers’ cognitive mechanisms, innovatively integrating a task-aware belief modeling module, dynamic observation-reflection mechanisms, and a causal explanation generation module. It significantly improves assessment accuracy and interpretability in complex contexts, enabling the automatic construction of structured cognitive diagnosis logic chains to provide scientific support for precise instructional interventions.

The research team has accumulated fruitful achievements in this field, including student knowledge state modeling frameworks and educational agent systems, with related works published in international authoritative journals such as *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, *IEEE Transactions on Industrial Informatics (TII)*, and *Knowledge-Based Systems (KBS)*, including highly cited papers and hot papers), as well as CCF Class A academic conferences like *AAAI* and *IJCAI*. As a key component of the

team’s cognitive intelligence technology innovation system, this study aims to tackle the core problem of dynamic contextual cognition modeling in next-generation intelligent educational systems, promoting the paradigm shift of educational assessment from static label diagnosis to dynamic cognitive reasoning.

This work was supported by the Regional Science Foundation of the National Natural Science Foundation of China (62567004), the National Natural Science Foundation of China (Young Scientists Fund, Class A) (62525606), the National Natural Science Foundation of China (Young Scientists Fund) (62406303), the Guizhou Provincial Science and Technology Projects (General Program) (QKHJC-MS[2026]753, QKHJC-MS[2026]754), and the University of Science and Technology of China-Liupanshui Normal University Counterpart Cooperation and Development Joint Fund (USTC-LPSNU-2025-10). This work was also supported by the Guizhou Provincial Science and Technology Projects (QKHJC[2024] youth012) and the Guizhou Provincial Higher Education Undergraduate Teaching Content and Curriculum System Reform Project (GZJG2024331, GZJG2024323).