基于多智能体的科学问题解决能力评测模型自动构建方法

Research on the Design and Application of Pair Programming Scripts for Primary School Students: Based on the Perspective of Productive Failure

陈小汾
江南大学人文学院，江苏无锡 214122
3174464009@qq.com

[摘要] 随着生成式大模型技术的快速发展，大模型驱动的多智能体在教育领域已获得广泛应用。将多智能体应用于科学问题解决能力的评测，有助于处理多模态数据、输出高效评测结果和提供个性化学习路径。本研究面向中小学科学教育场景，基于大模型的通用智能代理框架进行适配和改造，设计了科学问题解决能力评测的智能代理框架和多智能体代理自组织协同框架。通过实例示范，定义了智能代理的角色，展示了多智能代理的协同工作流，并详细描述了各代理的工作职责和流程。
[关键词] 科学问题解决能力；教育评测；智能代理

Abstract: *With the rapid development of generative large language model technology, multi-agent systems driven by large models have been widely applied in the field of education. Applying multi-agent systems to the evaluation of scientific problem-solving abilities helps to process multi-modal data, output efficient evaluation results, and provide personalized learning paths. This study, targeting primary and secondary school science education scenarios, adapted and modified a general intelligent agent framework based on large models, designing an intelligent agent framework for evaluating scientific problem-solving abilities and a multi-agent self-organizing collaborative framework. Through example demonstrations, the roles of intelligent agents are defined, the collaborative workflow of multiple intelligent agents is showcased, and the work responsibilities and processes of each agent are described in detail.*
Keywords: scientific problem-solving skills, educational assessment, AI agent

1.前言

科学问题解决能力是个体为实现某个目标，综合运用科学知识、探究技能和推理策略等，有计划地进行认知活动的能力。培养学生的科学问题解决能力已成为国际科学教育的核心目标，亦是我国科学素养培育的重要主题。2025 年 1 月，教育部颁布《中小学科学教育工作指南》（以下简称《工作指南》）中明确指出，需在探究实践中强化对学生问题提出、实验设计、动手操作、思维发展等能力的考察，重点评测学生利用知识、方法，分析问题和创造性解决问题的能力。然而，通过一线科学教师深入访谈发现，经验丰富的教师能够感知学生的问题解决能力，但这种感知往往属于默会知识，难以转化为有效的评测因素。学生在课堂中展现出的问题解决能力是可见可捕捉的，但如何将这些现象转化为可测量、可操作的实践，目前仍存在探索空间。

随着人工智能技术的日渐发展和广泛渗透到教育活动中，其有望为教育评测带来更高效、更客观的策略方案，以弥补传统评测手段的不足。《工作指南》强调，开拓生成式人工智能大模型在科学教育中应用的新场景，利用数据分析技术提升教学评测的精准化水平。2020 年 10 月，国务院印发的《深化新时代教育评测改革总体方案》中也提出充分利用信息技术提升教育评测的科学性、专业性、客观性。当前，智能代理（AI 智能代理）在教育评测应用中的潜力被越来越多的教育研究者和实践者广泛关注。智能代理被定义为通过

观察环境并利用其可用的工具对环境采取行动以尝试实现预设目标的应用程序。研究表明，基于大模型的单智能代理在实时分析学习数据、识别知识盲点、提供针对性练习和反馈等方面具有显著优势。然而，单智能代理在处理多角色协作的复杂教育任务时存在局限性，如认知视角单一、角色定位固定等，难以满足深层次的教育需求。

为解决上述问题，研究者们提出了大模型驱动的教育多智能体系统解决方案，旨在通过多个智能体的协作来突破单智能体的能力限制。多智能体系统通过整合不同角色定位的智能体，构建起一个动态协作的智能教学团队，能够激发出超越单智能体的集体智慧，有效应对复杂教育任务的挑战。实证研究进一步验证了多智能体系统的优势，翟雪松等人（2024）[xvii]基于微软 AutoGen 多智能体框架搭建人机交互平台，设计由七个智能体组成的教研团队作为“智能代理”角色，包括教师、学习者、教育心理学专家、教育社会学专家、教育经济学专家、教育评测专家、教育政策制定者，学生作为“人类”角色，与七个智能体角色基于问题开展集体研讨，结果表明，相比单智能体模式，多智能体能够显著提升学习者提问策略的多样性，有效提高复杂问题的解决效率。Yang 等人（2024）[xviii]提出 LLM 智能代理-CK 框架，由三种类型的 LLM 驱动的代理（即管理员、审判员、评论员）和两种控制策略（即讨论策略和决策策略）构成，通过智能体之间的协作与讨论，实现对教师数学内容知识的识别。Lagakis 等人（2024）[xix]利用大语言模型（LLMs）和 AutoGen 框架开发了一个基于多智能体架构的自动化评分系统，该系统包含管理员代理、评估代理和反馈代理，它们通过讨论和决策策略共同完成评分任务，被用于大规模在线开放课程（MOOCs）中的作业评分。

基于上述分析，本研究通过探讨多智能体系统赋能科学问题解决能力评测中的范式转变、内容创造、场景应用的可能性，构建基于大模型的科学问题解决能力多智能评测系统框架，并将其应用于科学问题解决能力评估实践，以期推动智能代理在教育评测中的创新应用，为科学问题解决能力评测提供技术路线与实践参考。

2.多智能体系统赋能评测范式转变

2.1 处理多模态评测数据

传统评测数据收集方法主要集中于文本形式的答题结果，这种单一数据形式难以全面捕捉分析被评测者在问题解决过程中的丰富信息。基于大模型的智能代理具有强大的理解与分析复杂语言结构的能力，已具备生成文本、图像、音频、视频及 3D 模型等多种模态内容的处理能力。在多智能体系统中，每个智能体能够独立感知并处理来自不同来源的数据，灵活调用外部工具来增强自身数据处理和分析能力，共享信息平台允许各智能体进行数据交换和同步，整合不同来源和模态的数据，共同协商形成一个统一的、全面的能力评测结果。

2.2 输出高效化评测结果

传统评测结果的输出需要花费较长时间进行批改和评分，无法及时给学生反馈，教师教学效果也无法实现实时评测，导致教学目标和流程的优化呈现滞后性。而多智能体系统可以实现实时评估，全程化高效率地进行过程性评测、增值性评测，通过即时生成的语言输出，快速给出反馈，让学生及时了解自己的不足之处，调整学习策略；帮助教师和相关部门及时了解课程的优势和不足，从而促进课程内容和教学方法的改进。

2.3 提供个性化学习路径

在学生评测中，多智能体系统基于大数据算法对每个学生的阶段性能力评测结果进行精准分析，定位学生当前科学问题解决能力所处水平。从发展性的角度将学生的总体科学

问题解决能力目标分解为具体的阶段性目标，识别学生在科学问题解决能力上的优势和需要改进的地方。通过分层分类的引导策略，系统帮助学生设定合理的阶段性学习目标，并根据这些目标提供定制化的学习内容和指导，促使学生在努力达成阶段性目标的过程中提高学习效率和成就感，实现评测从物到人、从外向内、从短期到长期、从结果向过程的转变。

3.基于大模型的通用智能代理框架

3.1 基本概念

在人工智能领域，智能体需要能够感知其外部环境并做出行动，以对外部环境产生影响。通常情况下，智能体与外部环境间的感知与行动不断循环，形成密切交互，以完成具体任务目标。图 1 为基于基于大模型的通用智能代理框架（Xi et al, 2025）[xx]，由大脑、感知、行动三个关键部分组成。其中，大脑又称为控制模块，由一个大型语言模型组成，不仅存储知识和记忆，还承担着信息处理和决策等功能，并可以呈现推理和规划的过程，能很好地应对未知任务。感知模块的核心目的是将智能代理的感知空间从纯文字领域扩展到包括文字、听觉和视觉模式在内的多模态领域。当一个智能代理拥有类似大脑的结构，具备知识、记忆、推理、规划和概括能力以及多模态感知能力时，它也有望拥有类似人类的各种行动来应对周围环境。在智能代理的构建过程中，行动模块负责接收大脑模块发送的行动序列，并执行与环境互动的行动。

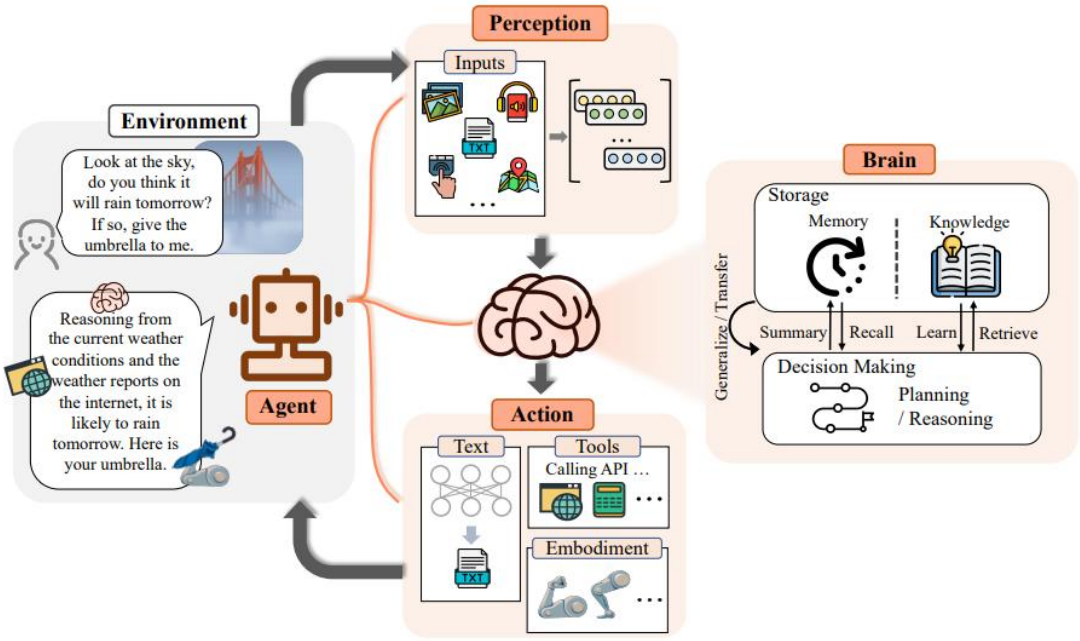

图 1 基于大模型的通用智能代理框架

多智能系统是由多个互相协作或竞争的自治智能体组成的系统，旨在通过集体行为解决复杂问题。系统中每个智能体基于通用智能体框架，分别配置感知模块、控制模块和行动模块，并通过独立或联合的训练机制实现功能优化，从而具备一定的自主性与适应性。如图 2，多智能代理协作框架的核心是如何实现智能代理之间的协作和竞争的平衡，即如何使每个智能代理都能达到自己的目标，同时也能促进整个系统的性能和效益。为了实现这一目标，多智能代理协作框架需要解决以下几个关键的问题:

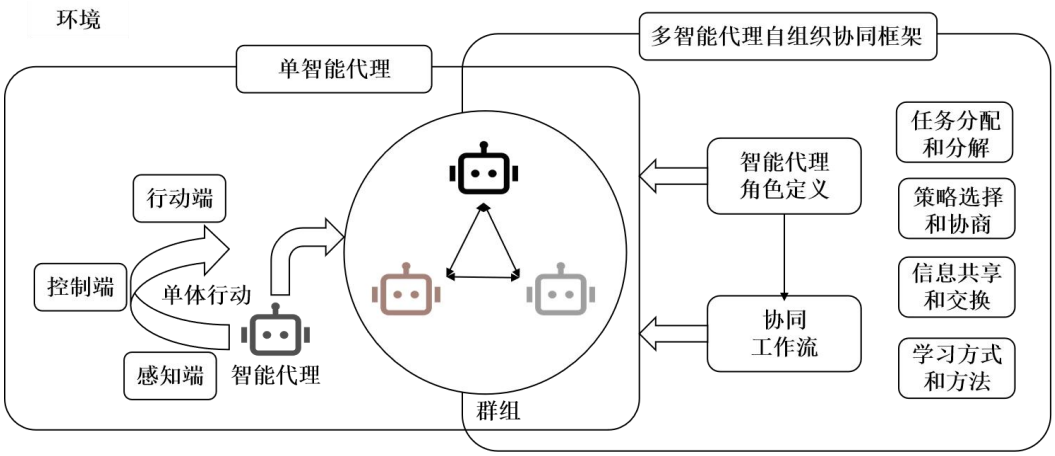

图 2 多智能体代理自组织协同框架

（1）智能代理的角色画像设计。定义智能代理的属性和行为，如目标、偏好、策略、动作、感知、学习和沟通等。角色定义采用多个单工具的智能体策略，针对任务目标各子任务，按照各要素子任务分配对应的工具，结合工具的文档说明对智能体进行训练，使智能体达到理解工具或独立调用接口等目标。

（2）任务分配和分解。将一个复杂的任务分配和分解给多个智能代理，使得每个智能代理都能承担合适的子任务，同时也能保证任务的完整性和一致性。任务分配阶段通过提示词进行任务的明确和分解，形成完成复杂任务的思维链，提取思维链中各序列化的节点作为子任务，形成任务序列清单。根据各智能体角色分工将各子任务分发下达给各智能体。

（3）动作处置和规划。将高层次的任务目标分解为可执行的低层次动作，并规划反思动作的执行顺序，通过环境感知反馈来进一步适应多层级搜索，以实现整体目标。在此阶段，多个智能体获得初始任务，通过多轮与环境的交互，优化智能体所分配的子任务和执行顺序，得到最优的规划。

（4）智能代理的协同工作流。设计智能代理之间的交互和协调的机制和协议，如定义工作流程、交互信息等。

4.科学问题解决能力评测的多智能代理框架

4.1 科学问题解决能力评测智能代理框架

针对图 1 所示的基于大模型的通用智能代理框架经适配与改造，用于科学问题解决能力的评测，如图 3 所示。

（1）感知模块。针对科学课堂活动中数据来源众多（研学单、实验操作过程记录、师生互动对话、小组讨论对话等）、数据类型模态多样（非结构化文本、图像、音频和视频流等）等问题，教师在学生科学问题解决能力评测中难以兼顾到每位学生，也难以从多源异构数据中提炼出一致的能力评估证据。感知模块将智能代理的感知空间，从纯文本拓展到包括文本、视频、图片和语音等多模态领域，使智能代理能够更有效地从复杂教育场景中获取与利用信息，突破多源异构数据信息识别抽取技术、多模态信息统一表征认知模型等关键技术，以支持对跨模态、多样式的要素信息的识别、抽取，多模态知识获取、表示与推理，形成多模态的统一语义表示，提升智能代理在多模态信息环境中的智能感知能力。

（2）控制模块。为处置评测任务进行处置动作规划，它可以决策处置动作需要运用哪些操作指令来完成，包括记忆、结构感知知识和处置动作规划 3 部分。记忆部分，存储和管理与科学问题解决相关的概念和知识、记录学生在实验中的具体操作和情境，保存学生在

解决问题过程中的思考路径和决策过程，进而支持智能代理的反思和进化。知识部分，通过动态图结构归纳自然语言文本和领域知识库的语义表示，强化智能代理的认知能力。包含评测框架、评估案例、科学学科核心概念集以及实验知识与规范操作，帮助智能代理深入理解学生的实验设计和操作是否符合科学标准。动作规划部分，根据记忆和结构感知知识，依托科学教育动作库和科学知识体系规划反馈，将复杂的评测任务分解为具体的子任务，并规划出最优的处置动作，生成评测报告和反馈建议。控制模块通过总结、召回、学习和检索等机制，将记忆模块、知识部分与处置动作规划协同工作。总结机制提炼记忆中的关键信息，形成结构化的知识表示，召回机制根据当前任务需求，提取相关的历史信息和案例，学习机制不断更新记忆和知识库，提升智能代理的认知能力和评测准确性，检索机制则确保智能代理能迅速获取所需信息，为当前任务提供及时支持。

（3）行动模块。负责将控制模块的规划结果转为具体行动，以实现对学生科学问题解决能力的智能评测和反馈，并输出包含文字和可视化图表的评测报告。包括检索器和自主调用模型；其中检索器根据输入的操作指令从数据集中检索与操作指令相关的接口，帮助模型过滤无关信息，提高模型的训练效率；自主调用模型，根据操作指令与接口进行多轮交互，并返回交互结果，通过教师与多智能体多轮对话中教师对评测结果提供的质疑进行修改和调整，动态评测报告内容。可视化模块，用于将收集到的学生表现数据以柱状图、折线图等方式直观展示，呈现学生在不同科学知识点或者技能点上的掌握程度以及在问题解决过程中的进步情况。行动模块在执行过程中需与感知模块和控制模块紧密协作，确保准确地接收感知模块处理后的结构化数据和关键信息，并按照控制模块的规划结果采取相应行动，同时将执行过程中的状态和结果及时反馈给控制模块，以便其根据实际情况调整规划策略。

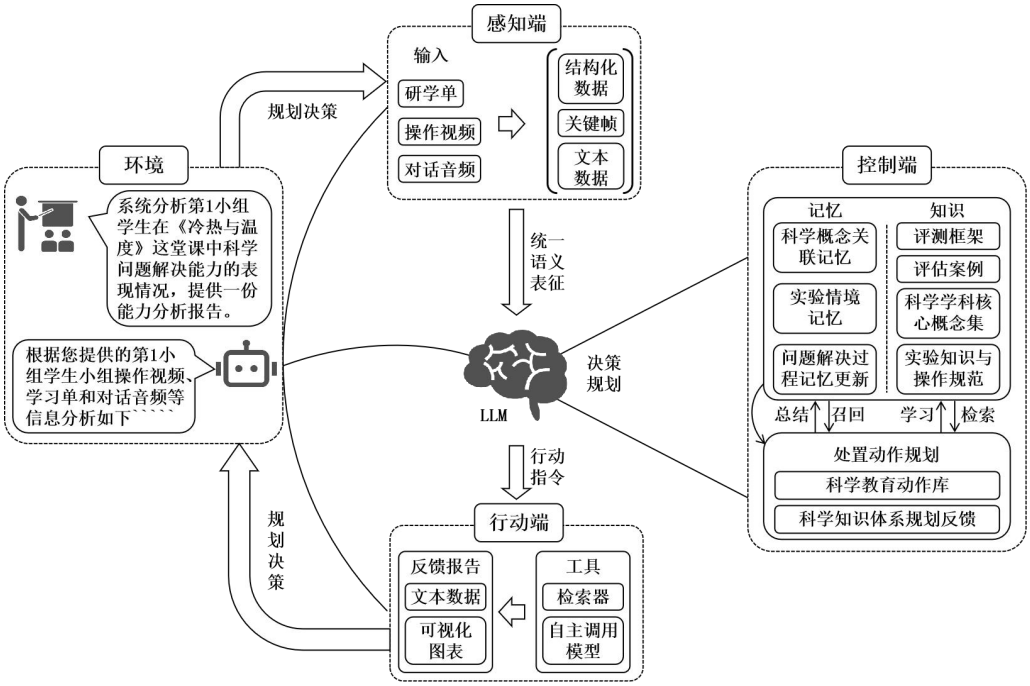

图 3 科学问题解决能力评测智能代理框架

4.2 科学解决能力多智能代理自组织协同评测运行框架

面向科学课堂中学生科学问题解决能力的评测，以图 3 科学问题解决能力评测智能代理框架为底本设计管理员和各模态分析员的任务和角色，其中，各模态分析员还需要进行差异化的训练，具体见表 1。

表 1 各模态分析员差异化设计训练

角色名称	感知模块	控制模块	行动模块
视频分析员	输入：操作视频流 技术：VideoPrism 模型提取时空特征，OpenPose 检测协作姿态 输出：操作步骤序列、异常动作标记	基于规则引擎（实验步骤校验）与深度学习（动作意图识别）混合架构 动态权重调整：根据任务类型分配检测重点	生成带时间戳的异常报告
语音分析员	输入：ASR 转写的讨论文本 技术：BERT+逻辑链检测模型 输出：假设提出次数、反驳语句占比	分层决策：进行初级过滤，去噪无关对话，后面在进行深度分析 集成动态记忆网络（DMN）记录讨论演进过程	生成讨论质量雷达图 触发知识图谱验证（如术语使用准确性）
文本分析员	输入：OCR 识别的研学单文本+手写公式 技术：结构化解析引擎 输出：数据完整性评分、结论逻辑错误	轻量级规则引擎，用以校验必填字段 与视频分析智能体协同验证时间一致性	输出结构化校验结果 标记需人工复核的模糊内容

见图 4，科学解决能力多智能代理自组织协同框架运行框架可划分为评测任务分解、评测数据自分析和评测结果整合四个实现阶段。首先，评测任务分解阶段，管理员智能体能够单独根据教师的作业意图和任务指令，理解当前学生所处具体科学问题情境，自动选择不同模态分析的分析员参与下一阶段。其次在评测数据自分析阶段，各模态分析员在在数据分析阶段，根据自动抽取与所负责数据的信息，通过多智能体分组讨论，执行评测框架中的能力维度映射任务，得到讨论结果和分析信息集合。最终在评测结果整合阶段，对讨论结果投票，将讨论结果映射为能力的高、中、低三类，并分别整合支持能力评测结果为高、中、低的三类分析信息，得到最终学生能力结果及其报告。

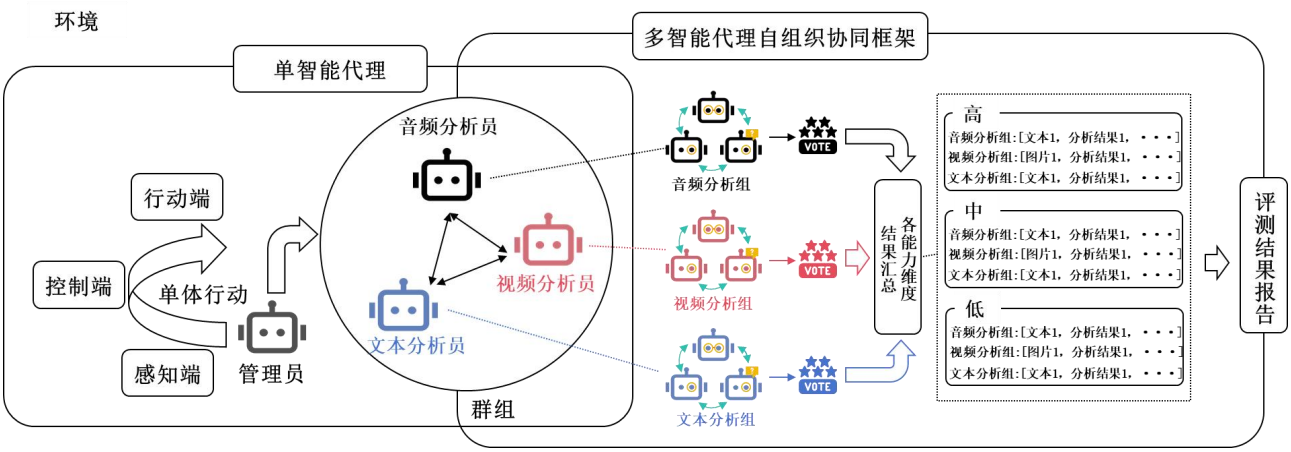

图 4 科学解决能力多智能代理自组织协同框架

（1）智能体角色画像设计

科学解决能力多智能代理自组织协同框架框架采用预先设置与模型生成相结合的方式设置智能体角色画像。该框架中每个模态讨论小组包括 3 个智能体，称为智能体 A、B、C。其中智能体 A 的角色画像通过大语言模型针对科学问题解决能力的评测框架生成；智能体 B、C 的角色画像设计根据讨论策略由人工给出。该方法在满足讨论策略需要的基础上，尽可能提高了智能体角色画像的多样性和专业性，以分化角色意见并提高知识调用的准确性。

5.各能力维度结果汇总

在评测讨论结果整合部分，要求全面考虑智能体间的不同观点，降低多智能体分组讨论过程中错误传播和智能体判断错误对最终分类结果造成的影响。

6.科学问题解决能力评测的多智能代理框架设计实现

为支持基于大模型的多智能体落地实现，目前已经有多个工程框架被开发并开源，例如 Auto Gen[xxi]、LangGraph[xxii]、MetaGPT[xxiii]、Google ADK[xxiv]等。这些框架为研究人员和开发者提供了重要资源，便于开发和测试多智能体的多场景应用。在上述实现框架中，Auto Gen 支持用户依据需求，灵活定义多智能体之间的交互模式与人机协作模式，例如由一位智能体主持、人工参与多智能体交互的动态群组讨论模式，或由两个智能体分别负责编码与执行调试的协作编码模式等。Auto Gen 可支持多智能体的交互记忆读写，可通过调用 Python 第三方工具包实现工具的使用（例如调用 Matplotlib 绘图库完成数学制图），并且支持将任务转化为机器语言解决，例如利用代码分步执行任务，并通过智能体间的代码执行与调试（Debugging）确保程序的成功运行。LangGraph 是 LangChain 生态系统中的一个重要工具，它提供了一个基于图（Graph）的框架来构建复杂的 LLM 应用。通过将应用逻辑组织成有向图 Q 的形式，LangGraph 可以让构建复杂的对话流程变得更加直观和灵活。

Auto Gen、LangGraph 等开发框架均为多智能体系统实现提供了可行方案，还可以通过配合使用发挥二者各自优势例如，可以利用 Auto Gen 灵活构建并实现智能体的交互框架与基于机器语言的任务执行，并利用 LangGraph 协助连接外部丰富的工具库（例如 Ar Xiv、Office365、Wolfram Alpha 等）以及自定义工具（通过用户提供工具功能描述、方法实现代码、输入输出格式等信息实现），从而拓展智能体的能力边界。

本研究选择基于 Auto Gen 和 LangGraph 框架进行评测工作流的搭建。如图 5，管理员在获取学生的多源数据后，进行深入分析，制定详细的组织处置方案。管理员将结构化的任务分配给各模态智能代理分析员，确保每个代理清楚自己的职责和任务。各智能代理根据分配的任务进行工作。在执行过程中，智能代理通过共享信息平台进行数据交换和状态同步，确保任务执行的连贯性和一致性。为了提升系统的适应性和响应能力，多智能体系统采用强化学习技术，为每个智能代理训练出独立的分布式决策策略网络，实现一定程度的自协同。此外，系统借鉴教育评测中的形成性评测理念，通过动态调整任务分配和执行顺序，不断优化工作流，确保评测过程的高效性和结果的可靠性。最终，各智能代理的执行结果被整合，最终输出生成详细的评测报告，为教师提供有力的支持，帮助他们更好地理解和指导学生的科学问题解决过程。具体工作流见图 4。

图 4《冷热与温度》课程第 1 小组科学问题解决能力的工作流示意图

7.结论

当下科学问题解决能力评测存在时效性差、分析周期长而影响教师对学生学习情况的及时掌握和教学策略的调整，亟需引入新技术、新方法、新框架推进科学问题解决能力智能分析。本研究基于大模型的通用智能代理框架，设计了科学问题解决能力评测的智能代

理框架和多智能体代理自组织协同框架。通过多智能体系统的协作，实现了对科学问题解决能力的自动评测，并提供了一个示范实例。

参考文献

翟雪松,季爽,焦丽珍,朱强 & 王丽英.(2024).基于多智能体的人机协同解决复杂学习问题实证研究.开放教育研究,30(03),63-73.doi:10.13966/j.cnki.kfjyyj.2024.03.007.

Yang, K., Chu, Y., Darwin, T., Han, A., Li, H., Wen, H., ... & Liu, H. (2024, July). Content knowledge identification with multi-agent large language models (LLMs). In International Conference on Artificial Intelligence in Education (pp. 284-292). Cham: Springer Nature Switzerland.

Lagakis, P., & Demetriadis, S. (2024, June). EvaAI: a multi-agent framework leveraging large language models for enhanced automated grading. In International Conference on Intelligent Tutoring Systems (pp. 378-385). Cham: Springer Nature Switzerland.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2025). The rise and potential of large language model based Agents: A survey. Science China Information Sciences, 68(2), 121101.

Yang, H., Yue, S., & He, Y. (2023). Auto-gpt for online decision making: Benchmarks and additional opinions. arXiv preprint arXiv:2306.02224.

Wang, J., & Duan, Z. (2024). IIntelligent Spark Agents: A Modular LangGraph Framework for Scalable, Visualized, and Enhanced Big Data Machine Learning Workflows. arXiv preprint arXiv:2412.01490.

Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., ... & Wu, C. (2023). Metagpt: Meta programming for multi-Agent collaborative framework. arXiv preprint arXiv:2308.00352, 3(4), 6.

Xu, H., & Shatz, S. M. (2003). Adk: An Agent development kit based on a formal design model for multi-智能代理 systems. Automated Software Engineering, 10, 337-365.