大语言模型赋能新工科实验教学的可靠性研究

王 婉　孔 梅
吉林建筑大学经济与管理学院，吉林 长春 130018

[摘要]在新工科实验教学加快数字化转型的背景下，大语言模型已进入资源组织、方案设计、操作指导和实验评价等环节，但生成内容的概率性、实验情境的复杂性以及教学评价的价值属性，使技术可用并不必然转化为教学可靠。本文运用文献分析、风险分析与系统设计方法，将可靠性界定为教学系统在规定情境和运行周期内，持续提供有据可依的知识支持、安全适度的学习帮助、公平可解释的评价反馈并保持教师有效控制的能力。围绕模型、任务、学习者与制度环境的交互关系，分析知识失真、情境错配、自动化依赖、评价偏移和版本漂移等风险的形成机理，提出由目标约束、证据支持、过程控制、评价校验和治理审计构成的可靠性保障框架。进一步从课前、课中、课后及跨课程运行四个环节给出实施路径，构建覆盖知识准确性、安全处置、过程稳定、评价一致性、学习迁移和可追溯性的验证思路。研究认为，大语言模型赋能实验教学的关键不是扩大自动决策范围，而是依据任务风险限定模型权限，以证据链约束生成，以人工复核守住安全与评价底线，并以学生脱离模型后的独立表现检验真实教学成效。

[关键词]大语言模型；新工科；实验教学；可靠性；人机协同；教学评价
DOI：10.33142/fme.v7i10.20703　　中图分类号：G642.0　　文献标识码：A

Reliability of Large Language Model-enabled Experimental Teaching in Emerging Engineering Education

WANG Wan, KONG Mei
School of Economics and Management, Jilin Jianzhu University, Changchun, Jilin, 130018, China

Abstract: As experimental teaching in Emerging Engineering Education undergoes digital transformation, large language models are increasingly used to organize resources, support design, guide operations, and assess learning. However, probabilistic generation, complex laboratory contexts, and the value-laden nature of educational assessment mean that technical availability does not necessarily produce instructional reliability. Using literature analysis, risk analysis, and system design, this study defines reliability as the capacity of an instructional system to provide evidence-based knowledge support, safe and appropriately bounded learning assistance, fair and explainable feedback, and effective teacher control under specified conditions and throughout its operation. It examines the mechanisms through which knowledge distortion, contextual mismatch, automation dependence, assessment bias, and model-version drift arise from interactions among models, tasks, learners, and institutional environments. A safeguarding framework is proposed, comprising goal constraints, evidence support, process control, assessment validation, and governance auditing. Implementation pathways are then developed for pre-class preparation, in-class activities, post-class assessment, and cross-course operation, together with verification dimensions covering factual accuracy, safety handling, process stability, assessment consistency, learning transfer, and traceability. The study argues that the central issue is not to expand automated decision-making, but to delimit model authority according to task risk, constrain generation through evidence chains, preserve human review for safety and assessment, and evaluate authentic learning through students' independent performance after model support is withdrawn.

Keywords: large language models; emerging engineering education; experimental teaching; reliability; human-AI collaboration; educational assessment

1 问题提出

新工科建设所面对的，不只是新知识进入课程的速度加快，更是工程问题边界、技术组合方式与人才能力结构的深层变化。相关政策强调学科交叉、产教融合和实践创新[1,2]，工程教育研究也持续讨论新范式、专业结构与培养模式的重构[3-5]。在这一转型中，实验教学承担着把抽

Copyright © 2026 by authors and Viser Technology Pte. Ltd.

象知识转化为工程行动的任务。学生既要理解原理，又要在设备约束、数据噪声和方案冲突中作出判断。实验因而不能只被看作验证结论的规定步骤，而应成为发现问题、比较方案和承担决策后果的学习过程。

大语言模型给这个过程带来了一个新的条件，大大提高了它的效率。可以利用自然语言来表述概念，按照学习者的提问重新组织信息资源，帮助解析代码和实验数据，给出有关报告中论点联系方面的评价意见。由于该类实验课程存在师资力量欠缺、实验室设备种类繁多、学生基础参差不齐等状况，所以它有很强的实际吸引力。一条看似语法正确但是出处不明的操作指南，会把微小的知识错误变成设备的安全隐患，同样的，一篇结构完整的但是隐藏着评分标准误解的自动化评估反馈也不是绝对可靠的。如果一个模型的回答具有很高的自然流畅性，那么使用者就会忽略它的生成过程以及它所依赖的证据边界，这就体现出设计和使用这类系统的平衡技术性能和透明度的要求。

现有讨论较多关注大语言模型能做什么，对其在何种条件下值得信任、出现错误后如何纠正以及长期运行是否稳定，仍缺少面向实验教学全过程的分析。教育研究已经指出隐私、偏见、透明度、过度依赖和教师责任等问题[6,7]，自然语言生成领域的幻觉研究也表明，文本连贯不能代替事实正确[8]。但如果只把可靠性缩减为模型准确率，又会遗漏教学目标、学生发展和组织治理等关键维度。即使每次回答均有事实依据，系统若直接替代学生推理，仍可能损害能力形成；即使模型版本没有变化，设备手册更新、任务情境改变和评价标准调整，也可能使原有回答失去适用性。

根据以上观点，本文主要从单个功能效果评价的角度出发，转而对教育系统的稳定性展开研究，并试图回答三个主要的问题，即怎样衡量大语言模型对于新型工程学科实验教学的支持程度是可靠的？主要的风险是如何沿着一定的途径产生并流传出去的？高校应当创建起怎样的体系来保证它的运作流程有执行力、检验性以及问责制，进而有效地维持系统运转？文将模型置于课程目标、操作规程与教师职责构成的制度框架中，为课程设计、平台建设与质量评定提供依据。

2 可靠性的理论界定与分析框架

2.1 从模型性能到教学系统可靠性

工程学上可靠性是针对某个对象在规定的环境和时间内完成预定的功能所进行的评价。将此思想运用到实验教学当中，首先需要认识到对象不是单一的一个大语言模型，它是一个复杂的、综合的教学生态系统，这个系统包含着课程内容、实验活动、模型应用、知识库、教育者、参与者、管理机制等诸多方面。通用测试成绩只能体现模型对于某一个领域基础能力的掌握情况，并不能由此得出该模型对于某个学校设施、专业场景或者某个学生群体来说是否合适有效。可靠性须在具体学习结果、风险任务、证据体系与责任归属中界定。

实验教学中由于不断有新的技术、新的设备、新的教学方法出现，所以完全消除所有的错误是很难做到的。更为重要的是该系统有没有发现不确定性，能否有效地阻止局部错误进入高风险的环节当中，又是否可以保存足够的记录来支持责任追溯和之后的改正工作。可靠性包含预防、监控、应对和恢复这四个方面。

2.2 教学可靠性的四个维度

第一，知识可靠性，即模型所提供的概念解释、操作依据、数据处理方法与课程审核资料一致，关键结论能够追溯到有效来源。检索增强生成把参数化模型与外部资料检索结合起来，为知识密集型任务提供了重要思路。知识可靠性还应包括来源时效、文档版本、上下文完整性以及对证据缺口的明确提示。对于安全规程、设备限值等高风险内容，规则和标准应高于模型生成，不能用语言推断填补制度空白。

第二，对于过程可靠性的主要方面来说，就是指系统在全部的教学过程中，能够持续而有效地执行预先设定好的教学方案。当模型解析仪表图像的时候产生误读，根据这些错误的信息给出调整建议的时候，会造成设备运行状况明显变坏的情况发生。过程可靠性包含平台在线稳定的同时也牵涉到输入验证、各个阶段的操作权限控制、系统状态实时同步、人工干预等各方面的因素。

第三，评价可靠性，即系统对不同学生、不同时间和不同表达形式的同类学习表现作出一致且有依据的判断，自动分析不替代教师作出高风险评价。评价并非纯粹的技术识别，而是关于学习成果的价值判断。测量效度理论提醒人们，分数解释与使用必须有证据支持。据此，自动评价不能只追求与教师分数表面接近，还应检查它是否真正测量了课程目标，是否对非标准表达、小组分工和特殊学习需要形成不当惩罚。

第四，发展可靠性，即智能支持没有以短期任务完成替代长期能力成长，学生在撤去帮助后仍能独立分析、规范操作和迁移方法。支架的价值在于帮助学习者跨越暂时无法独立完成的困难，并随能力增长逐步撤除[9]。如果系统始终给出完整方案、代码和结论，实验报告可能更加整齐，学生却失去经历不确定性与修正错误的机会。发展可

Copyright © 2026 by authors and Viser Technology Pte. Ltd.

靠性由此成为连接技术质量与教育质量的关键维度，也是判断“赋能”是否成立的最终依据。

2.3 任务、证据、行为、结果分析链条

本文用任务、证据、行为、结果这样的分析框架来贯穿全文。从任务角度出发，要准确识别出学习目标以及可能存在的风险等级，进而确定哪些部分可以由模型自己产生，哪些部分只能通过检索来呈现，哪些部分必须由教师亲自参与管理。证据维度就是对输入的数据是否可靠、知识来源是否权威、评价材料是否完整等进行评价，保证决策有理有据。行为维度主要研究模型给出建议、学生执行操作、教师干预三者之间如何配合，以达到改进教学过程和学习体验的目的。结果维度是从任务完成的速度、系统的运行状况和学生自主能力的提高等方面考虑，对整个教育活动进行评价，看它是否有效果、有没有产出。四个相互联系的方面一起起着制约的作用，使得系统评价不是依靠一个输出或者一个满意度指标来决定的，而应该从整体上考虑它的性能好坏。

分布式认知理论认为，认知活动分布于个体、工具、符号和环境之中[10]。在这一视角下，大语言模型已经成为实验认知环境的一部分，其可靠性不能脱离人机分工讨论。合理分工不是把教师工作尽可能交给模型，而是把规则清晰、可重复核验的事务交由技术辅助，把情境权衡、安全处置、价值判断和最终评价保留给教师。学生也不是被动接受建议的对象，而应参与证据核查、方案比较和错误申诉，形成对智能工具的必要审视。

3 大语言模型赋能实验教学的主要可靠性风险

3.1 知识失真与实验情境错配

大语言模型利用概率来生成文本的时候，会碰到产生虚假信息、模糊参数设定和不恰当地迁移同类设备的操作指南到现有设备上的问题。实验技术具有很强的情境依赖性特点，即对于相同的实验仪器而言，由于型号不同、量程不同、固件版本的不同都会造成操作规范存在较大差别，同样的数据处理算法在不同的样本特性或者理论假设下也会失去有效性。如果通用模型缺少对课程知识库以及设备详细的元数据的支持，那么它所给出的答案就存在着表面上看起来是正确的但实际上隐藏着危险的特点。

情境错配常常是由于问题表述不完整造成的。学生给出部分图像或者只说数据错误的回答，就会造成模型没有考虑实验目的、连线环境和前置操作，就直接指向最常见的一种解决办法，而忽略问题的复杂性与多样性。如果系统没有指明要增加哪些核心细节，那么这样的建议很可能是建立在未经证实的假设之上的。

3.2 自动化依赖与学习能力空心化

实验学习过程重视学生的亲身参与，即设计方案、动手做、反思改正等环节，用直接的经验加深认识和掌握技能。当模型可以很好地产生完整的解题步骤的时候，学生就会把注意力集中在模仿答案上，并且美化自己的报告，从而忽略对这些变量之间相互影响以及它们所支持的证据进行详细的探究。此类依赖性在初期一般会表现出工作效率明显提高的情况，这种状况容易造成课程管理者对于它可能存在的危险没有充分的认识。

在教育领域里，教师群体也存在着对自动化技术产生依赖性的倾向，这体现出的是技术应用同教育实践相融合过程中出现的一种复杂的动态。当教师面对大量系统自动生成的学习状态标识、评价标准的时候，如果因为时间紧促等缘故而直接使用这些数据的话，那么教学决策就变成了一个被隐藏起来的算法规则的附庸，而不是依靠教师丰富的专业知识以及敏锐的直觉所做出的选择。为了防止学生过分地依赖自动化系统，在教育实践活动中应该采取如下一些措施来加以防范，即明确免模型操作的适用范围和限制条件，提倡学生主动参与，用口头或者书面的形式表述自己的决策过程和思考逻辑，教师要实行定期的检查制度，将自动化的反馈同实际的学习效果进行比较，从而达到对学生理解程度和运用能力的全面考察目的。

3.3 过程证据失真与评价偏移

智能实验平台可以收集到对话交互、用户点击行为、方案不同版本的迭代、代码执行过程以及生成的报告文本等各种各样的数据，但是丰富的数据并不一定就可靠有效。如果把这类容易得到的数据当作衡量学习投入的标准的话，那么就会偏向于评价那些已经熟练掌握了平台操作方法或者擅长口头表达的学生，而忽略了其他一些同样重要的学习能力或者贡献。模型对于文本的处理上对于风格、语言习惯、表面一致性的特点比较敏感，会造成即使内容实质一样但是表达方式不同的作业得到不同的评分结果。

更深层次的隐患是生成的内容会破坏证据链的完整性和真实性。学生用模型建构策略加自我反思的时候，教育系统采用同样的模型去识别它的内在能力特点，就存在着一种危险，那就是系统会误把模型本身具有的表达形式当作学生的真实理解程度。如果评估过程中没有包含口头答辩、现场操作或者独立迁移任务这些重要的部分，那么评价体系就会陷入对文本质量的循环论证当中，而无法真实地反映出被评价者的能力状况。

3.4 版本漂移与多环节耦合失效

大语言模型的服务会不断升级，提示模板、检索手段、

Copyright © 2026 by authors and Viser Technology Pte. Ltd.

知识库和外部接口这些部分都会做相应的改善和改变。任何一个改变都会造成相同的输入得到不同的输出结果。版本漂移现象在长期课程管理当中更加明显，它体现出一个道理，即便以前的教学过程经过证实是有效的并且被批准通过了，也不能保证在之后的版本里就可以随意地继续使用，因为系统的更新或者软件的升级都会使得原先的流程变得不可靠起来。

如果生成、检索、设备识别、评价这四个模块之间存在任何一个环节上的偏差，都会造成耦合失效的情况出现，这就需要引起注意了。因为检索过程中出现了遗漏的情况，使得模型证据集不够充分，模型依然用确定的语言回答问题，评价系统之后就把这个答案当成正确答案来记录下来，于是就形成了一个表面看起来完整但实际上却是错误的反馈循环。

4 面向全过程的可靠性保障体系

4.1 以学习产出和任务风险限定模型权限

可靠性设计首先应回答什么不许自动化，而不是什么都能自动化。课程团队可依据成果导向教育理念，将毕业要求转化为可观察的实验表现，再反向确定任务、支持和证据[11]。CDIO 所强调的构思、设计、实现与运行过程，可用于识别各阶段的真实学习需要[12]，避免模型因技术上能够实现就越过教学边界。

任务风险一般分为低级、中级和高级三种层次，这样划分有利于对各种任务的不确定因素及可能产生的影响进行系统的评价和控制。就概念阐述和资料查找这类低风险工作而言，可以依靠模型的辅助功能，在需要的时候给出具体的出处引导。对于方案对比分析、数据校验等中等风险的任务，应该严格按照实证依据来进行操作，同时要保证学生参与到决策过程中来，并且通过教师的随机抽查来监督执行的质量。设备安全联锁配置、危险参数阈值设定、最终评价结果的得出，只能由制度或教师决定，模型不得直接干预实际操作。

4.2 建设可审计的知识底座与证据链

建立可信知识库要以教师审阅过的资料为主，包含大量的、重要的实验指导书、仪器使用说明书、安全操作规程、典型故障分析、数据处理规范和评价指标体系等。每一份文档都要有清楚的标记，标明它对应的课程名称、使用的设备型号、软件版本、有效期限以及审核人等，对于已经失效的文件要采取相应的隔离手段。在执行检索的时候，应该先找到和当前任务以及特定设备有关的信息，而不是依靠网络上的通用经验去代替学校的内部操作规程。

系统要创建起一个针对关键建议的最少必要记录，该记录包含学生所提出的问题，相关背景资料，引用的文本片段，模型给出的回答，规则检验结果以及教师所做的修改意见等内容，从而达到对决策过程核心部分有完整地把握的目的。

4.3 建立分级提示、双重校验与人工接管机制

面对学生第一次遇到难题的时候，教学体系应该先让学生认真检查任务要求以及现有的观察结果，然后慢慢加入一些相关的概念提示或者类比例子，在确实需要的时候再给出全部的解题途径。在给学生提供提示之后，要细心安排阶段让学生去解释或者推测，以此来保证学生思考的过程不会马上指向答案，进而促使学生有更深层次的理解和运用。伴随着学生能力的发展，系统的步骤提示越来越少，取而代之的是用开放式的提问、促进知识迁移的方式来激发学生主动探究、深刻理解。

关键输出要经由双重验证来保证它的正确性以及合规性，该机制包含两部分，即知识来源和语义一致性检查通道以及依据具体规定进行操作流程、参数设置和禁止事项审核的通道。当两种分析途径得出不同的结论的时候，应该按照安全规程以及教师的指引来进行。遇到数据超出范围、缺少重要信息、连续出现非正常情况或者模型不能给出合理的解释等情况的时候，系统就会停止自动提案的过程，把任务转交给人类教师去处理。

4.4 以多源证据校验评价可靠性

智能评价体系应该把课程标准当作核心来创建，而不是仅仅依照模型容易识别的文本特性去安排。对于客观要素，比如字段完整性检验、代码测试、数据格式是否符合规范等都可以用预先设定好的算法来进行自动评价，但是涉及语义上的部分，例如论证结构是否合理、各种方案之间的比较分析等，则更适合用模型生成初步的标签来帮助理解。创新性成果、工程实践中所遇到的伦理问题、团队成员具体分工及合作效率、最后的成绩评定等都要依靠教师全面考虑之后做出判断。在评价模型性能的时候，输出的结果要准确对应到具体的评价指标项以及它背后的数据证据上，不能只是给出一个模糊的评分等级。学生有权对自动化的即时反馈提出质疑，教师在纠正这类反馈之后，应当把它加入之后的校准案例当中，从而保证系统是准确无误并且公平公正的。

校验评估要从三个方面考虑，即一致性、公平性和有效性。一致性检验就是对不同时段、不同提示、不同模型版本下对同一个任务的评价结果是否一致进行检验。公平性检验主要是探究，在给定相同的文本、使用不同的语言

Copyright © 2026 by authors and Viser Technology Pte. Ltd.

风格、允许出现合理的非标准答案的时候，由于没有教学依据而产生的偏见所造成的评价上的差别。效度检验就是把自动化评定的结果同实际的操作表现，口头答辩的成绩，迁移运用的能力等联系起来加以考察，从而判定评价系统是否可靠有效。当自动评分系统同人工评价产生持续性的差别之时，第一项工作应当是对评价标准及其所依仗的证据根基展开仔细的审视，而不是仅仅期望教师去顺应已有的技术平台。

4.5 形成覆盖模型全生命周期的治理审计

可靠性治理需要明确专业学院、实验中心、信息化部门和教师的职责。专业学院审定学习目标与评价规则，实验中心确认设备资料和安全边界，信息化部门负责访问控制、日志保护与运行监测，任课教师承担课堂处置和最终评价。供应商可以提供技术服务，但不能替代学校决定教育标准。生成式人工智能教育应用所涉及的伦理和人的主体性问题，也要求制度设计保持透明与可问责。

模型、知识库、提示模板和量规均应独立编号，更新前完成固定案例与边界案例测试，更新后保留回退方案。审计不应只检查是否发生事故，还要定期抽查证据引用、人工接管、学生申诉和教师修改，识别尚未造成后果的薄弱点。数据采集坚持教学必需原则，明确保存期限与访问范围，避免以可靠性为名无限扩大个人数据收集。风险管理框架所强调的治理、测量与持续改进，可为这一过程提供方法参照。

5 可靠性保障在实验教学环节中的实施

5.1 以课程实例完成准入验证

在课程正式进入课堂之前，课程开发团队要选择能体现该课程主要特点的任务来创建一个典型的测试集，而不是直接用供应商给定的示范案例。为了对系统是否适合于各个领域进行全面评价，测试集需要经过精心的设计，包含各种各样的标准问题和复杂的状况，如参照旧版手册、非标准设备型号、模糊图像、缺关键信息的问题及可能诱发越权的异常输入。

对于学生的前置自学活动主要是对学生已有知识进行评价，确定学习过程中操作的范围。系统会根据概念测验结果给出个性化的推荐资源，但是不能过早地制定出全面的实验方案。学生进入实验室前要经过安全确认程序，清楚哪些具体的事项需要向指导老师汇报。教育实践当中，教育者依照诊断评价结果来灵活调节学生分组方式以及任务难易程度，目的在于防止把模型化的要求当作固定的学程引导，进而推动个性化的教学契合与发展。

5.2 把安全控制与独立思考嵌入交互

课程实施是以实验任务清单为驱动，模型只是获取它在当前阶段需要的数据信息。对于学生提交的异常情况，系统先要检验设备状态、操作步骤和数据的有效范围，之后再给出相应的解决办法；对存在风险的操作只显示经过批准的应急停止和上报程序。教育管理系统中教师界面应该清楚地列出待处理的事件以及它们的优先级，保证重要的事情最先被注意到，不会因为一些琐碎的对话信息而被掩盖掉。针对网络中断、识别错误或者响应超时等情况，实验过程中设置了 fallback 机制，即当出现以上情况的时候可以自动切换到不需要依靠某个模型的其他流程上，从而保证教学活动和安全控制不会因为一个技术平台的故障而受到影响。

为了有效地防止学生出现抄袭的行为，在完成重要的步骤之前应该先做一次预测工作，即确定要观察的现象是什么，用什么理论来解释它，如何判断分析方法是否正确。完成操作之后，要仔细比较预测结果同实际观测数据，剖析二者之间存在的差别以及产生这些差别的原因。

5.3 以答辩、复核和申诉闭合评价链条

通过整合实验报告、操作日志、方案迭代版本和反思文档来完成分类证据，并找出其中存在的疏漏问题，教育工作者主要任务是评价这些资料的教学意义以及启发作用。就某一个学生而言，在答辩中应该把重点放在选择方案、分析出现的异常情况、运用方法论等几个方面，而不是简单地重复别人的研究成果。课程结束不等于可靠性验证结束。教师还应安排撤除模型支持的小型迁移任务，观察学生能否在新装置、新数据或新约束下独立完成关键步骤。如果报告质量提高而迁移表现没有改善，就需要重新审视提示是否过强、任务是否过于模板化。学习分析可以支持理解与决策，但指标解释必须回到学习产出，不能把平台活跃度本身当作改革成效。

5.4 以回归测试支持持续改进

当把教育体系由一门课程扩展到多门课程的时候，不能用同一个指导方针和评价标准来对待不同的课程及其目的。各个专业领域要按照各自特有的设备风险特点、成果展现形式、前置学习要求来重新确定并具体化相应的权限设置及评价标准。

每当模型或者知识库被更新的时候，应该先做一次安全基线以及高频任务的回归测试，保证它们是稳定的并且有效的，然后再有条不紊地把它们恢复到正常使用当中去。就评价体系来说，为保证其稳定可靠，需要保存一份相同

Copyright © 2026 by authors and Viser Technology Pte. Ltd.

的匿名作业作为参照物来考察各个版本之间是否存在非正当变化或者不确定性的波动。当核心指标出现下降的时候，可以采取回退到以前稳定版本或者限制应用规模等手段来保证系统的性能同预期的目标相一致。

6 可靠性评价指标与验证设计

6.1 建立技术、教学与治理相结合的指标体系

为了对系统可靠性进行全面评价，应该采用多角度检验的方法来保证各个主要指标之间互相印证，从而达到系统可靠性的目的。就知识层面而言，主要是看事实是否正确，并且是否有权威性的引用，信息来源是否及时更新，在缺少证据的时候如何谨慎地对待；就过程来说就是观察系统对于异常状况做出反应的速度、作出决定的精确度以及恢复速度等各方面的情况；就评价方面而言，是从人机交互判断的一致性、不同的版本之间的稳定性、系统对于各种表达形式的适应性以及有效的申诉和修正程序等方面来进行评价；就发展而言，是研究系统在没有先验条件约束的情况下可以自主学习的能力、口头讲解的清晰程度、新任务中迁移的效率；就治理而言，注重版本管理全过程可追溯、日志完整、访问权限有效控制、整改措施落实到位，创建起闭环式管理机制。

这些评价标准不能变成一个综合评分。当安全基准没有达到要求的时候，即使系统平均的知识准确度很好也不能说明它的运行是可靠的。为了保证分析结果的准确性、可比性，在给出指标结果的时候要详细列出测试环境介绍、样本特征统计、使用的模型具体版本等信息，防止脱离某种情况的数据解读。对于目前缺少足够的证据来支持的指标应该明确标示出来作为观察对象，不能用估计值代替真实的测量结果。

6.2 采用分层测试与故障注入检验系统

技术验证过程可以分为三个层次，即单元验证、链路验证和情境验证。单元测试主要针对检索、生成、规则校验、日志记录这些基本的功能模块做详细的验证工作，以保证它们是正确的、有效的；链路测试就是对系统的各个模块之间相互配合的稳定性进行检测，用模拟的方式给系统输入过期的规程等操作来检验它能否正确地阻止错误的蔓延；情境测试采用师生参与的方式进行实际操作，在控制的环境里完成具体的任务，对提示机制的顺畅程度、信息传达的效果以及系统面对异常状况的处理能力展开全方位的考察，从而保证复杂的操作环境下系统表现出的效果同预期一致。

故障注入机制要覆盖各种可能的问题情景，即知识库为空的时候、网络通信慢的时候、图像数据不清楚的时候、传感器工作不正常的时候、学生互动策略出错的时候、模型版本更新的时候等等各种情况。对于评价模块设计来说，应该加入等值改写、非标准但是合理的解决办法以及包含常见错误类型报告等内容，从而达到对反馈机制稳定性和与既定量规一致性进行全方位检测的目的。所有发现的问题都要形成可以重复再现的修复案例，在之后的更新中防止这类问题再次出现。

6.3 以准实验和纵向证据检验教学成效

教学验证可在条件相近的班级采用分阶段实施或准实验设计，比较不同支持策略下的学习表现，但不把班级成绩差异简单归因于模型。前测用于了解先修基础，过程记录用于分析提示使用与错误修正，后测关注近期学习，延迟迁移任务则考察独立能力。教师工作记录、学生访谈和异常事件材料可补充量化结果，帮助判断变化来自任务设计、教师投入还是模型支持。早期人工智能教育综述曾提醒教师视角不足[13]，这一缺项在可靠性实证中应得到补足。

研究应报告负面与无差异结果，不只呈现成功案例。如果某种教育技术手段虽然可以减少重复性问题解答的需求，但是却会增加教师的复查工作量，那么我们就应该做成本效益分析，客观地评价它的整体影响。若措施使部分学生受益却加重薄弱学生对教师的依赖，应调整支持策略，而非只看平均效果。可靠性检验要依靠长时间的数据积累，在不同的学生群体中、各种任务的情境之下、不断更新的模型架构里开展系统的试验并加以评判。

7 讨论

7.1 可靠性治理的教学价值

本文架构将技术效能、教育实践与制度义务纳入连续评价过程，重塑智能实验教学的决策与责任分担。可靠性治理机制对加深认识教育工作者的角色有着十分重要的作用。学生已经由原来的单纯的内容接受者变成了证据的验证者，他们的工作就是对所接受或者拒绝的建议提出自己的看法。这种人机交互模式把批判性评价变成日常的操作，同在一个实验任务框架之下同时培养起人工智能意识和技术工程能力。

7.2 适用边界与研究局限

不同的实验课程对于可靠性的要求是不一样的。在虚拟仿真环境下，错误修正一般可以通过低成本的回滚来完成，但是涉及到真实的高压、高温或者化学实验的时候，就需要采取更加严格的权限隔离措施来保证安全。对于程序设计实验来说，用自动化测试机制可以得到清楚准确的反馈信息，但是对于开放性工程设计项目来说，由于它的复杂性以及多变性，很难用一个标准去评价设计方案的好坏。本文建立起来的框架主要是给出一套设计原则和检验方法的思想，目的是帮助各个专业的人员按照一定的规范

Copyright © 2026 by authors and Viser Technology Pte. Ltd.

来进行安全性的论证工作，该框架并不是一种直接代替各个专业内部具体分析和评价的过程。

本文主要是用理论分析和系统的构想方法，目前没有根据某个课程的长期教学效果去估计出关键指标的临界值，也没有依靠实践的经验来判定教育成果的好坏。所提出的知识、过程、评价和发展这四个主要方面，在各种模型结构以及课程安排的应用情境之中，它们的重要性以及影响程度还需要经过实证研究加以证实。后续研究工作建议选择一门实验性质的课程开展全面周期的试点项目，目的是探寻并公开化地制订测试集准则、人工介入操作日志以及版本更新办法，从而以此为根基展开对于可靠性加强措施同教师工作负担、管理开销以及学生自主学习能力长久发展之间联系的评价。

8 结论

大语言模型对于新工科实验教学的应用已经远远超出了仅仅检验答案是否正确这样简单的评价范围，它变成了一个包含知识资源整合是否有效、实验操作是否高效、学习成效是否公平、教育机构管理机制能否协调运转在内的庞大系统。本文根据任务、证据、行为和结果之间内在联系，对知识失真、情境错配、自动化依赖、评价偏差、版本漂移这些可能的风险进行了分析并加以区别。在此基础上创建起一个包含目标约束、证据支撑、过程控制、效果检验和治理体系审计这五个部分相互渗透的全方位保障体系，从而有效地对上述各种风险进行管理并加以降低，保证决策的过程是正确的、可靠的。其主要机理就是用风险等级划分确定模型访问权限，依靠可追溯证据规范核心产出过程，由教育专家保证安全和评估标准的坚持，实行全面的生命周期测试保证系统能长期稳定运行。

实验教学的主要目的就是培养学生独立思考的能力，在遇到工程上不确定的因素的时候，可以做出正确的、合理的判断，而不是为了得到答案而流利地表达出来。为了对大语言模型赋能的效果进行全面评价，不仅要考察系统是否可以做到准确、稳定地运行，还要考察学生在辅助支持移除之后，能否有效地说出自己的决策依据、按照标准程序来操作，并且能够把学到的方法应用到新的情境中去。技术指标与学习效果互证、效率改进受安全公平约束时，大语言模型才能成为持续推动实验教学变革的动力。

基金项目：2025 年度吉林省教育科学规划一般课题：《多智能体驱动的新工科人才培养实验教学体系改革与实践》（GH25094）。

[参考文献]

[1]教育部高等教育司.关于开展新工科研究与实践的通知[Z].教高司函〔2017〕6 号,2017.

[2]教育部,工业和信息化部,中国工程院.关于加快建设发展新工科实施卓越工程师教育培养计划2.0的意见[Z]. 教高〔2018〕3 号,2018.

[3]吴爱华,侯永峰,杨秋波,等.加快发展和建设新工科 主动适应和引领新经济[J].高等工程教育研究,2017(1):1-9.

[4]顾佩华.新工科与新范式：概念、框架和实施路径[J].高等工程教育研究,2017(6):1-13.

[5]林健.面向未来的中国新工科建设[J].清华大学教育研究,2017,38(2):26-35.

[6] KASNECI E, SESSLER K, KÜCHEMANN S, et al. ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and Individual Differences,2023(103):102274.

[7] YAN L, SHA L, ZHAO L, et al. Practical and ethical challenges of large language models in education: A systematic scoping review[J]. British Journal of Educational Technology,2024,55(1):90-112.

[8] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys,2023,55(12):1-38.

[9] WOOD D, BRUNER J S, ROSS G. The role of tutoring in problem solving[J]. Journal of Child Psychology and Psychiatry,1976,17(2):89-100.

[10] HUTCHINS E. Cognition in the Wild[M]. Cambridge, MA: MIT Press,1995.

[11]顾佩华,胡文龙,林鹏,等.基于“学习产出”（OBE）的工程教育模式——汕头大学的实践与探索[J].高等工程教育研究,2014(1):27-37.

[12] CRAWLEY E F, MALMQVIST J, ÖSTLUND S, et al. Rethinking Engineering Education: The CDIO Approach[M]. 2nd ed. Cham: Springer,2014.

[13] ZAWACKI-RICHTER O, MAR Í N V I, BOND M, et al. Systematic review of research on artificial intelligence applications in higher education — where are the educators?[J]. International Journal of Educational Technology in Higher Education,2019(16):39.

作者简介：王婉（1983—），女，汉族，吉林长春人，博士，副教授，吉林建筑大学经济与管理学院，研究方向为工程管理智能决策；*通信作者：孔梅（2001—）女，汉族，山东曲阜人，硕士在读，吉林建筑大学经济与管理学院，研究方向：建设工程管理。

Copyright © 2026 by authors and Viser Technology Pte. Ltd.