人工智能通识入门微课 Part 3 人工智能辅助评估

Neo Lee

<2025-06-10 Tue>

目录

评估是教育的指南针，引领教学与学习过程不断进步、迈向卓越。

人工智能工具在评估领域同样可以帮助我们提高质量和效率。不过应该特别注意以下几点：

- 人工智能工具可能使评估过程机械化或包含特定偏见，可能忽视人类的主观判断和情境因素，我们应明确认识到人工智能评估不能完全替代人类评估，最好结合起来以保持评估的灵活性和全面性；
- 现阶段人工智能生成的内容依赖于其训练数据，而训练数据并不能保证都是“安全”的，我们应时刻关注人工智能生成的内容，防范其可能的内容风险及知识产权风险；
- 如果用于正式的、重大的阶段性评估，应格外关注评估过程及结果的隐私信息保护，避免无授权的人或系统访问到敏感信息；
- 评估的结果往往对学生带有较强的影响和引导力量，所以也带来更强的主体责任，我们应明确认识到现阶段的人工智能只是工具，最终对一切行为结果负责的还是作为主体的、使用工具的人。

这一部分我们将通过实践案例来学习与体会人工智能工具在评估领域可以带来的帮助及启发。由于评估的目标及方法与具体学科强相关，下面将分别展示数学和英语两个学科的案例。

1 3-1 数学学科的 AI 评估

1.1 目标与问题

学习运用人工智能工具实施一次对数学教学中特定单元或知识点的评估。

更进一步，希望深入了解和理解：

- 有哪些人工智能工具对此有帮助?
- 使用人工智能工具的正确流程和方法是怎样的?
- 使用过程中应避免哪些误区?

1.2 初步研究与分析

我们可以向自己常用的大语言模型提出类似这样的问题：

> 结合实际操作案例，说明运用 AI 工具辅助中学数学学科设计和实施测评的能力，以及正确使用方法
>
> **提示**
>
> - 类似这样比较复杂和综合的课题，最好选用比较先进、并带有深度思考能力的大语言模型，比如 DeepSeek R1（在 DeepSeek 应用中开启“深度思考”开关），或者通义千问、豆包等模型的类似模式；
> - 如果一个模型的答复不能令人满意，可以多试几个，并将多个结果作为输入要求模型对其继续进行反思与提炼。

通过这样的迷你研究项目，可以帮助我们快速凝聚相关领域知识、整理和提炼我们自身的经验与思考，结合起来就更容易得到正确的尝试路径。

1.3 实践案例

1.3.1 目标场景

初中三年级，刚学完函数图像的平移、伸缩、对称变换。教师需要快速了解学生对不同变换规则（如“左加右减”、“上加下减”、系数影响等）的掌握程度，找出普遍性错误和个别薄弱点。

1.3.2 传统流程

教师手动从题库选题或编题，耗时费力，题目类型和难度覆盖可能不够全面或精准。

1.3.3 AI 辅助流程

1. 教师明确需求 (核心)

   教师清晰定义测评目标：

- 知识点：诊断学生对基本函数（如 $y=x^2$, $y=1/x$ 等）进行平移、伸缩（仅限沿 y 轴）、对称（关于 x 轴、y 轴、原点）变换的理解和应用能力；
- 题目难度梯度：基础识别 → 单一变换应用 → 组合变换应用 → 逆向求原函数，需要包含常见错误选项（如混淆平移方向、伸缩系数影响）。

2. AI 工具选择与输入

   选用具有数学内容生成能力的先进大语言模型，如 DeepSeek 或通义千问，或专门定制的学科模型，如 九章 或 MathGPT。

   根据上述评估需求，设计如下提示词（prompt）输入：

   > 你是一位经验丰富的初中数学教师。请为我生成一份针对“函数图像变换（平移、沿 y 轴伸缩、对称）”的诊断性小测验，包含 10 道选择题和 2 道简答题。
   >
   > 选择题：覆盖所有变换类型（平移、伸缩、对称），包含常见错误选项（如将“左移 2 单位”误操作为“减 2”）。难度递进。
   >
   > 简答题：1 道要求画出给定函数（如 $y=2(x-1)^2+3$ ）的图像并描述变换过程；1 道给出变换后的图像和部分信息，要求逆向推导原函数或其表达式。
   >
   > 请确保题目表述清晰、严谨，符合初中数学课程标准。最后请提供标准答案和详细的评分标准（特别是简答题的步骤分）。

3. AI 生成初稿

   AI 根据提示生成测验初稿，包括题目、选项、答案和评分建议。

4. 教师审阅与修改 (核心)

   教师是质量把控者。教师仔细检查：

   - 准确性：题目本身是否有科学错误？选项是否合理？答案是否正确？

- 适切性：难度是否符合班级水平？表述是否清晰无歧义？是否覆盖了所有关键知识点？错误选项是否真的反映了常见误解？
- 教学意图匹配：是否完全符合本次诊断的目标？是否需要调整或补充？
- 评分标准：是否合理？步骤分划分是否科学？
- 格式：排版是否清晰易读？

教师根据审查结果进行必要的修改、删减或补充题目。

5. 实施测评

   教师在课堂或在线平台发放修改后的测验。

6. AI 辅助初步分析 (可选)

   如果使用的在线测评平台支持或集成有 AI 分析功能，通常可以实现：

   - 自动批改选择题；
   - 快速统计每题正确率、错误选项分布；
   - 识别出错误率异常高的题目和选项。

7. 教师深度分析与反馈 (核心)

   教师结合 AI 提供的初步数据：

   - 聚焦问题：查看错误率高的题目，分析学生具体错在哪里（如：普遍选了某个错误选项），判断是概念性错误还是操作失误；
   - 关注个体：查看每位学生的答题情况（尤其简答题），识别个别学生的特殊困难；
   - 制定策略：基于分析结果，决定下一步教学重点（是全班巩固某个概念，还是分组辅导，或对个别学生进行单独指导），并在课堂上进行有针对性的讲评和反馈。

1.3.4 AI 工具体现的能力

- 快速生成能力：大幅缩短教师命题时间
- 多样化题目生成：能生成不同类型（选择、简答）、不同难度的题目，并尝试模拟常见错误
- 初步数据分析：结合平台快速提供客观数据，帮助教师聚焦问题

1.3.5 小结

- 教师主导需求定义：AI 是执行者，教师是需求定义者和决策者
- 严格的质量审查：绝不能直接使用 AI 生成的原始内容，教师专业审查是保证测评效度和信度的核心
- AI 作为效率工具：节省命题和初步批改统计的时间，让教师有更多精力进行深度分析和个性化反馈
- 结合具体平台功能：选择和用好在线测评平台的数据分析能力

1.4 拓展练习

换一个知识点，自行设计并完成一次类似上面的评估流程。

对比一下通用大语言模型和针对数学学科专门训练的模型之间的效果差异; 尝试一下九章大模型和 MathGPT 等专用工具提供的一些更专门的功能，比如各种九章的各种智能助手，MathGPT 提供的直接生成评估卷的功能。

2 3-2 英语学科的 AI 评估

2.1 目标与问题

学习运用人工智能工具实施一次对英语教学中特定能力或知识点的评估。

除了数学学科实践中探索的问题以外，我们应该意识到英语学科测评的难点在于：数学侧重逻辑与计算，而语言学习更注重语言应用能力（听说读写）和文化理解。因此在本案例中最好能关注三个新的维度：

- AI 如何解决语言类测评特有的主观题批改（如作文）问题；
- 如何创设真实语境；
- 对文化要素的评估。

2.2 初步研究与分析

类似 3-1，略。

2.3 实践案例

2.3.1 目标场景

某单元学习主题涉及“科技发展与环境挑战”。希望设计一份阅读理解练习，既能诊断学生不同层次的阅读能力（细节查找、推理判断、主旨归纳、词汇语境义、批判性思考），又能根据学生答题情况提供即时、个性化的反馈和强化练习建议。

2.3.2 传统方式

教师选取一篇难度适中的文章，设计统一题目。批改后统一讲解，难以针对个体差异提供即时、精准的反馈和补救措施。

2.3.3 AI 辅助流程

1. 教师明确目标与分层需求

    - 核心目标：评估并提升学生在主题相关文本中的多层次阅读理解能力；

- 分层需求：题目需清晰对应不同认知层次（如：L1 细节查找，L2 推理判断，L3 主旨归纳/作者观点，L4 词汇语境义/长难句分析，L5 批判性思考/联系现实）；
- 个性化反馈需求：学生答错后，能立刻获得该题考查能力的解释和针对性练习建议。

2. AI 工具选择与输入

选用擅长文本处理和题目生成的先进大语言模型，如 DeepSeek 或通义千问，或专门定制的相关学科模型，如 QuillBot 或 ReadTheory。

根据上述评估需求，设计如下提示词（prompt）输入：

> 你是一位高中英语教学专家。请基于以下要求，为我设计一份关于“科技发展与环境挑战”主题的分层阅读理解测评任务：
>
> - 文本选择：提供一篇适合高一学生（CEFR B1-B2 水平）的英文文章（约 400-600 词），主题聚焦科技（如 AI、可再生能源）对环境的影响（积极/消极/争议）。文章需包含学术词汇、复合句、不同观点。
> - 题目设计 (5-7 题)：
>   - 明确标注每道题考查的能力层次（L1 到 L5）；
>   - 题型：选择题为主（便于即时反馈），可包含 1 道主旨归纳简答题；
>   - 覆盖：细节事实 (L1)、隐含意义推理 (L2)、主旨/作者态度 (L3)、词汇/句子在语境中的含义 (L4)、开放性问题如“你同意作者观点吗？为什么？”(L5)；
>   - 为每道选择题提供高质量的错误选项（反映常见误解，如过度推断、偷换概念、忽略否定词）。
> - 即时反馈设计 (重点)：为每道选择题设计：
>   - 答对反馈：简短肯定，点明考查的能力点，如“很好！你准确地找到了支持该观点的细节信息”；

- 答错反馈：必须包含：
  - 能力点解释：清晰说明该题主要考查什么能力，如“本题考查的是根据上下文推断作者未明说的态度”;
  - 错误原因分析：针对错误选项，解释为什么它是错的，如“选项 B 看似相关，但它过度推断，文中并没有明确证据支持这个极端结论”;
  - 针对性微练习建议 (可选但推荐)：提供 1-2 个非常简短的练习建议或思考提示，如“回顾文章第三段，找出暗示作者对 XX 技术担忧的关键词”或“尝试用自己的话总结作者在最后一段的主要论点”。
- 提供标准答案和详细的答案解析，特别是推理题和词汇题。

3. AI 生成初稿

   AI 生成文章、分层题目、选项、反馈语、答案及解析。

4. 教师审阅与修改 (核心)

   - 文本审查：语言难度是否合适？主题是否符合？观点是否平衡？有无错误或偏见？
   - 题目审查：每道题是否准确对应标注的层次？题目和选项是否严谨无歧义？错误选项是否确实有干扰性且反映典型错误？L5 开放题是否有启发性？
   - 反馈语审查（重点）：AI 生成的反馈语往往过于笼统或机械，我们需修改使其：
     - 具体明确：直指考查的能力点和错误根源。
     - 建设性：提供可操作的改进建议或思考路径。
     - 鼓励性：即使答错，语气也要积极，引导学生反思。
     - 语言适切：符合学生理解水平。

- 答案解析审查：确保推理过程清晰、词汇解释准确。

5. 实施测评

方式一在线平台

- 将审阅修改后的内容导入支持即时反馈的在线学习平台(如 ClassIn, Moodle, Quizlet 或其他专门的阅读平台)；
- 学生在平台上完成阅读和答题；
- 系统根据答案自动推送对应的反馈语和微练习建议。

方式二课堂纸质材料 + 教师引导

- 分发纸质材料，学生答题；
- 学生答题完成后，教师公布答案；
- 学生根据错题，参考教师预先打印好的、对应题号的"反馈与建议卡"进行自我反思；
- 教师随后集中讲解共性难题。

6. 教师数据分析与跟进

- 平台自动汇总各题正确率、各能力层次（L1-L5）的整体表现；
- 教师查看个体报告：哪些学生在哪个层次上频繁出错？谁在 L5 批判性问题上表现出色？等等；
- 制定策略：
  - 全班层面：重点讲解错误率高的能力点（如推理题技巧）；
  - 分组/个体层面：根据 AI 反馈建议和教师观察布置不同的强化练习，如给推理弱的学生额外 L2 推理题，给词汇弱的学生相关语境词汇练习，引导 L5 强的学生进行延伸讨论或写作等。

2.3.4 AI 工具体现的能力

- 高效生成主题文本与分层题目；
- 初步构建错误选项和反馈框架；
- 自动化推送个性化反馈。

2.3.5 小结

- 分层设计是关键：教师必须清晰定义能力层次要求，AI 才能据此生成；
- 反馈语的质量是灵魂：AI 生成的反馈是“毛坯”，教师必须深度加工，使其具体、有指导意义、人性化，避免使用生硬的模板化反馈；
- 数据驱动教学：利用 AI/平台提供的分层数据，实现精准教学干预；
- 教师是反馈的“调音师”和干预的“设计师”：AI 提供即时反馈和初步数据，教师负责解读数据、设计后续深度学习活动。

2.4 拓展练习

换一个评估目标，自行设计并完成一次类似上面的评估流程。