

建设可复用的临床推理教学与评估平台：从理论整合到LLM驱动的个性化学习

摘要

临床推理能力是医学教育的核心目标，但现有的教学和评估工具往往各自为政——虚拟患者平台、智能辅导系统、LLM评分工具分别解决了“完整流程训练”“多维度分析”“自动评分”的问题，却鲜少在一个统一的框架下整合这些能力。本项目并不试图再造一个全新的教学系统，而是基于现有的虚拟患者、智能辅导和LLM评估等研究成果，做一次**系统性的整合**，并在尚未被“做满”的关键空白上进行**工程化创新**——统一临床推理框架、规范LLM评估工作流，并把多源数据转化为可用的个体化学习反馈。这项工作的核心包括三个相互支撑的部分：

1. **临床推理教学框架**：提出基于理论证据和教学实践的“六步诊断推理 + 双维评估维度”的统一框架，融合CRI-HT-S、SNAPPS、IDEA等已有工具的优势，同时纳入LLM在推理任务中的失败模式分析。
2. **LLM-as-Judge评估工作流**：设计一套标准化、可复用的评分工作流（从信息提取→对标对比→多维打分→理由生成），配套多模型合议与置信度路由机制，确保评分的可靠性与工程可实施性。
3. **学生个性化评估与建议**：把多维、纵向的数据聚合为可读、可行动的“推理画像”——包括推理风格定性、强项/短板地图、具体改进建议——让评估最终落在学生的真实学习需求上。

这三个部分形成一个闭环：理论框架为评估提供结构，评估工作流产生数据和理由，数据聚合为个性化画像和建议。本报告分别阐述这三个部分的设计逻辑、实现细节和与现有工作的关系，并在最后讨论平台的演进路线和局限。

第一部分：动机与现状分析

1. 背景：已有的基础设施已相当完善

临床推理是医学实践和医学教育中公认的核心能力，其重要性已在大量文献中得到论证。过去十多年，医学教育界在“用数字化工具系统训练临床推理”这件事上已取得显著进展，形成了三条清晰的主线：

1.1 虚拟患者与情境化教学平台

心血管虚拟患者e-learning平台、Body Interact以及一系列虚拟患者系统，已能提供从病史采集、体格检查、辅助检查选择、鉴别诊断、最终诊断到治疗方案、并发症和预后的完整流程训练。这类平台在“还原真实情境、覆盖完整诊疗流程”方面已相当成熟，有的还引入多模态数据（影像、听诊音频等），甚至结合社交机器人或LLM虚拟患者角色增强沉浸感。[Awada2024][BodyInteract2025]

1.2 智能辅导与多维度表现分析

Hepius、COMET等早期智能辅导系统通过NLP或Bayesian学生模型，对学生的病史提问质量、诊断术语使用、知识掌握水平进行自动化分析。Alteach进一步发展，提出严谨性（Rigor）、逻辑性（Logic）、系统性（Systematic）、敏捷性（Agility）、拓展性（Expansion）五个量化维度，并通过雷达图和纵向曲线展示学生在多个病例上的进步，同时提供错误诊断日志帮助教师识别班级常见误诊模式。[Hepius2021][Alteach2022]这些系统共同说明：把临床推理拆成多个可测维度，并在时间轴上跟踪，是可行且有教育价值的。

1.3 大模型 (LLM) 进入评估和推理任务

近期工作将LLM引入医学教育评估。在虚拟患者对话中，用LLM打分学生的临床推理质量（信息采集完整性、推理链条的逻辑性等），与专家评分对标。在Script Concordance Test中，用多个LLM组成“虚拟专家组”，为每个选项生成概率分布并对学生作答评分，同时生成自然语言反馈。在OSCE场景，用Whisper转录2000+场真实考试视频，再让GPT-4等模型对“是否进行了病史总结”这一沟通子项评分，输出结构化评分（statement extracted、rationale、score），与人类评分的Cohen's κ 达到0.88；当多个模型一致时， κ 进一步提升至0.95。[VPDialogue2025][SCT2025][OSCE2024] LLM在这里不再只是“做题选手”，而是逐渐扮演“评分者与反馈生成者”的角色。

2. 现状：功能分散，整合不足

现有平台在“教学流程、自动评分、学生分析”三大维度上的覆盖情况存在明显的不均衡。通过对16篇主要文献和系统的系统性对标，我们发现：

系统/研究	线索获取	线索解释	假设生成	假设评估	多维评分	纵向追踪	LLM置信度	干预建议
DDxTutor	X	X	✓	✓	X	X	X	X
Hepius	✓	✓	✓	✓	✓	X	X	X
Alteach	✓	✓	✓	✓	✓	✓	X	X
心血管VP	✓	✓	✓	✓	✓	X	X	X
Body Interact+LLM	✓	✓	✓	✓	✓	X	X	X
混合VP系统	✓	✓	✓	✓	✓	✓	X	~
社交机器人VP	✓	✓	✓	✓	✓	X	X	~
SCT+LLM面板	~	~	✓	✓	✓	✓	✓	~
OSCE LLM评分	✓	~	~	~	✓	X	✓	X
VP对话LLM评分	✓	✓	✓	✓	✓	X	X	X
医疗模拟AI辅导	✓	✓	✓	✓	✓	X	X	✓
ITS综述	~	~	~	~	~	~	X	✓
EHR推理	✓	✓	✓	✓	✓	X	X	X
AI临床教练	✓	✓	✓	✓	~	X	X	✓
COMET群体辅导	~	~	✓	✓	✓	~	X	✓
香港大学AI模拟患者	✓	✓	✓	✓	~	~	X	~

这份对标清晰地显示：这些能力往往存在于不同系统、不同论文之中，很少被整合进一个“以学生为中心、以学习评估为核心目标”的统一平台。特别是，关键维度如“LLM置信度”和“干预建议”，几乎完全缺失。

3. 缺口：整合、标准化与个性化

在承认现有平台已覆盖大量关键能力的基础上，我们仍能明确看到几个尚未被“做满”的空白：

Gap 1 – 缺乏统一的理论框架

虽然CRI-HT-S、Revised-IDEA、SNAPPS等工具在教育学层面已相当成熟，但它们通常围绕单一环节或表现形式设计：CRI-HT-S专注病史采集、Revised-IDEA专注文书推理、SCT专注“给新线索后的概率调整”。当前各个平台通常只采用其中一部分，很少在一个统一系统中整合多个工具，构建覆盖整个推理过程的“评估链”。

Gap 2 – LLM评分缺乏工程规范

关于“如何用LLM打分”已有许多成功案例，但都偏“一次性实验”性质：每篇论文单独设计prompt、规则和输出格式，聚焦某个具体子任务。尽管个别工作（如OSCE多模型合议、SCT LLM panel）已实践了结构化的输出和多模型合议策略，但这些方案还没有沉淀为一个可在不同任务节点之间复用的评估模块接口，也缺少在教学平台上持续地判断评分confidence并据此路由（自动接受/标记/推人复核）的工程指南。

Gap 3 – 评估缺乏个性化深度与可执行性

Alteach已展示了雷达图、学习曲线、错误诊断日志等做法，但几乎没有系统进一步走完后半段路径：用纵向多维数据生成学生的自然语言个人报告，不只是告诉学生“你在某维度得60分”，而是解释“在呼吸系统病例中，你的信息采集较完整，但在缩小DDx时经常忽略X类诊断”；对学生的推理风格进行定性分类（直觉型 vs 分析型 vs 混合型）；在报告中给出具体可执行的建议而非泛泛而谈。现有平台多停留在“把分数和图画出来”，而把“解释、诊断学习问题、给出行动建议”交回给教师。

小结：从缺口到创新方向

基于这些gap，本项目通过填补这些空白来对医学教育平台做进一步优化。具体而言，我们在**保持对现有研究尊重**的基础上，针对性地：

1. 统一了理论基础（第二部分）——不是否定CRI-HT-S或IDEA，而是把它们和LLM失败模式分析整合成一个从“线索获取”到“反思校准”的完整框架；
2. 规范化了LLM评估流程（第三部分）——总结出可复用的四步操作和置信度驱动的人机协作路由，而不是停留在零散的prompt engineering实验；
3. 打通了数据到决策的全链路（第四部分）——从多维评分数据自动生成可理解、可行动的学生报告，填补现有平台“只有指标、没有解释和建议”的空白。

接下来，让我们详细阐述这三个创新方向。

第二部分：核心方向一 - 临床推理教学框架

4. 临床推理的六步拆解与二元评估维度

在本项目的理论框架中，我们基于假说演绎模型（Hypothetico-Deductive Model）、问题表示理论（Problem Representation）与多个医学教育系统的实证，将诊断性临床推理拆解为六个层级递进的步骤。每一步既可独立训练和评估，也可动态循环。

4.1 病例框架化与情境设定 (Case Framing & Context)

定义：医生在接触患者前，需要明确“我现在在哪个临床场景、我要解决什么决策问题”。

理论基础：Yazdani等（2017）的综述指出，全科医疗与专科医疗、初诊与随访、急诊与门诊场景中，疾病谱、线索稀缺度、诊断目标各不相同。例如在全科，医生往往不需要给出终极诊断，只需判断“是否需要转诊或进一步检查”。Nature Digital Medicine的最新研究强调，医疗AI系统的设计应与实际临床工作流对齐。[Yazdani2017] [Gaber2025]

教学意义：让学生显式陈述“当前场景”（例如：“患者在急诊科初诊，需要快速判断是否生命危险”），而不是直接跳到列DDx。这一步的表现反映学生对临床现实的理解程度。

评估维度（从临床推理文献中总结）：

- **场景识别准确性**：学生能否准确识别病例的临床背景（初诊/随访、急诊/门诊等）
 - **诊断目标明确性**：学生能否清晰陈述“在这个场景中我要回答什么问题”
-

4.2 线索获取 (Cue Acquisition)

定义：通过问诊、体格检查、初步实验室检查等手段，系统性地收集关于患者的关键信息。

理论基础：Elstein等经典的假说演绎模型（1978）将临床推理分为四个成分：cue acquisition → hypothesis generation → cue interpretation → hypothesis evaluation，形成循环。这说明线索获取不是一次性的完整收集，而是与假设生成、假设评估动态交织的过程。[Elstein1978] Alteach等系统从真实医学记录中提取关键问题，评估学生是否能问到足够的关键线索。[Alteach2022]

实证研究表明，医生实际上是边获取边形成假设的，而非先完整地问历史再列DDx。因此，这一步的训练应强调“策略”而非“记忆”——学生需要学会在有限时间和资源约束下，优先获取哪些信息。

评估维度（来自CRI-HistoryTaking Scale与临床推理教学）：

- **信息完整性**：是否收集了足够的关键临床信息（症状时间轴、诱因、伴随症状、既往史等）
 - **信息质量**：所获取的信息是否明确、无冗余
 - **信息组织能力**：是否能将零散信息组织成“病史故事线”
-

4.3 问题表征 (Problem Representation)

定义：将收集的零散线索整合成高信息密度的精简表述，包括主症状、时间特征、严重程度以及语义修饰符（如“急性”vs“慢性”）。

理论基础：Mayo Clinic的“Exercises in Clinical Reasoning”详细演示了问题表征的形成过程：医生通过语义修饰符（semantic qualifiers）对病情进行分类编码。例如，“急性进行性胸痛伴呼吸困难”的表征比长篇症状罗列更能有效地触发诊断假设。[Regehr2018] 问题表征的质量与最终诊断成功率高度相关。

根据这些研究，我们可以看到问题表征是一项关键的思维过程中介步骤，介于“原始数据收集”和“诊断假设生成”之间。

评估维度：

- **信息密度**：Summary statement是否抓住主诉、时间、关键症状和病程演变，每条信息都有诊断意义

- **术语准确性**：是否使用医学术语（“急性/慢性”“进展性/间断性”等）准确描述，而非患者原始表述
 - **信息对齐**：最终的问题表征是否包含了之前收集阶段的所有关键信息，是否有遗漏
-

4.4 假设生成与初步排序 (Hypothesis / Differential Diagnosis Generation)

定义：基于问题表征，生成合理的诊断假设列表（鉴别诊断），并做出初步排序。

理论基础：Elstein模型的关键发现是**假设生成是早期事件**，医生在获取仅几个线索后就开始列初始DDx。进一步，**假设的质量（而非数量）**与诊断成功高度相关。[Elstein1978] DDxTutor等系统将推理拆分为“局部推理”（每条线索对各DDx的支持程度）和“全局推理”（整体DDx列表的合理性与排序）。[DDxTutor2025]

最新LLM在诊断中的研究强调，需要评价DDx的**广度**（是否涵盖所有关键候选）与**排序合理性**（是否与病例特征、流行病学匹配）。[Nori2025][Qiu2025]

根据这些研究，我们总结出这一步的核心理论工具：**区分“生成了多少诊断”与“这些诊断排序是否合理”两个独立维度**，避免“列了很多诊断就是好”的误区。

评估维度：

- **诊断覆盖度 (Breadth)**：是否包含所有关键的、常见的可能诊断，特别是高危、易遗漏的诊断（如主动脉夹层、肺栓塞）
 - **诊断合理性 (Appropriateness)**：是否避免明显不符合情境的诊断
 - **排序合理性 (Ranking Logic)**：初始排序是否与病例特点、流行病学概率一致
-

4.5 假设评估与信息收缩 (Hypothesis Evaluation & Narrowing)

定义：基于新的临床线索或诊断测试结果，更新对各假设的信念，逐步收缩DDx列表，最终做出诊断或管理建议。

理论基础：原发性全科医疗诊断策略模型中的“细化阶段”描述了医生在这一步使用的多种策略：限制性排除规则、逐步细化、概率推理、临床预测规则。[Yazdani2017] Script Concordance Test的教学原理是给出新线索，让学生判断每个诊断的概率如何变化，模拟现实中“在信息逐步揭示过程中的贝叶斯推理”。[SCT2025]

这一步的核心是“**诊断增益**”（某个新检查能改变诊断后验概率的程度），帮助学生区分“有信息增量的检查”vs“低效的乱查”，培养资源意识。

评估维度：

- **收缩的合理性**：是否把不符合新线索的诊断思想上降权或排除，而非机械地保留整个列表
 - **关键鉴别点识别**：是否能识别出区分不同诊断的关键特征（例如RA vs OA的关节累及模式）
 - **检查/信息的针对性**：为缩小DDx选择的检查/新线索是否真正能改变后验概率，是否有战略性
-

4.6 反思与校准 (Reflection & Calibration)

定义：在给出诊断/决策后，对整个推理过程进行元认知检查，识别可能的偏差、遗漏或过度自信。

理论基础：“Exercises in Clinical Reasoning”中强调“Time-out and Reflect”策略。eLife综述“Critique of impure reason”系统性地区分了“推理结论”与“推理过程本身”，强调仅看最终答案对不对是不够的。[Regehr2018]

[Sim2025] Nature Digital Medicine的PRM (Process Supervision Reward Models) 通过在每个推理中间步骤的自我反思与修正，改进了LLM的最终诊断准确性。[Zhang2025]

根据这些研究，我们看到反思能力是专家型医学推理的标志性特征，也是防止认知偏差的关键机制。

评估维度：

- **偏差识别能力**：能否识别"what doesn't fit"的线索并主动调整假设
 - **自信度评估**：能否评估自己诊断的可信度水平（高 vs 低 vs 不确定）
 - **条件性校准**：是否能进行"假设诊断"的反思（"如果X检查结果出来，我的诊断会推翻"）
-

5. 二元评估维度：过程导向 vs 结果导向

在实际教学中，需要两条平行的评估线：一条评估推理过程的健全性（过程导向），一条评估诊断结论的准确与安全（结果导向）。

5.1 过程导向：推理迹象评估（Rationale-Based Evaluation）

核心思想：不仅看"你的答案对不对"，还要审视"你是怎么想的"。

CLEVER (Clinical LLM Evaluation by Expert Review) Rubric的核心维度包括Factuality (事实正确性)、Clinical Relevance (临床相关性)、Conciseness (简洁性)。[CLEVER2025] Nature Communications研究统计了每个 reasoning step的正确率，发现有时中间步骤比最终答案更能反映推理质量。[Qiu2025] IDEA Assessment Tool 评价"Interpretive summary""Diagnostic reasoning rationale""Decision-making basis"等具体的推理迹象。[IDEA2014]

评估维度：

- **步骤覆盖度与完整性**：各个推理环节是否都有阐述
- **线索-假设链接的合理性**：从线索到假设的推导是否合理，是否存在逻辑跳跃
- **事实准确性**：提及的临床特征、风险因素等是否符合当前医学知识
- **表达简洁度**：是否信息密集而无冗余，避免啰嗦
- **反思与自我校准能力**：是否能识别"what doesn't fit"并调整假设，是否有元认知检查

5.2 结果导向：结论评估（Conclusion-Based Evaluation）

核心思想：无论推理过程如何，最后的诊断对不对、安全不安全、是否全面。

LLMEval-Med基准采用"可用性评分"，强调0-5打分中 ≥ 4 才算"在临幊上可用"。[LLMEvalMed2025] Nature 的"Towards accurate differential diagnosis with large language models"直接评价Top-1/Top-k诊断准确度以及列表的合理性。[Nori2025] 大型语言模型用于疾病诊断的范围审查强调了典型的结果指标：准确性、安全性、过度/欠缺诊断率等。[Scoping2024]

评估维度：

- **诊断准确性**：Top-1 accuracy (排首位的诊断是否正确) 和Top-N accuracy (前N个中是否包含正确答案)
- **诊断列表的合理与全面性**（Appropriateness & Comprehensiveness）：是否避免离奇诊断，是否遗漏高概率/高风险诊断
- **管理/建议的安全性**（Safety）：给出的诊断和管理建议有没有明显危险漏洞

- 资源使用与效率 (Efficiency)：提议的检查/管理步骤是否经济合理
 - 临床交流质量 (Clinical Communication Quality)：表达是否符合临床规范，对患者/家属的解释是否清晰
-

小结：从框架到评估工作流

第四和第五节阐述的六步框架与双维评估维度，为接下来的LLM评估工作流提供了明确的“评什么”和“怎么评”的答案。简言之：

- 六步框架定义了临床推理的结构化步骤，每一步都有具体的认知目标和评估维度；
 - 双维评估确保了既看“思路对不对”（过程），也看“结论对不对”（结果）；
 - 这两个理论基础为第三部分的LLM-as-Judge工作流提供了可操作化、可分化的评估模板。
-

第三部分：核心方向二 - LLM评估工作流

6. 原子化的LLM-as-Judge评估方法

在医学教育中，LLM-as-Judge的核心工作流可抽象为四个关键操作。每个操作都是原子化的，可以在任意推理步骤和任意评估维度上复用。

6.1 信息提取与标准化 (Information Extraction & Normalization)

目的：将学生的原始、自由文本回答转化为机器可读、结构化的“可评估陈述”。

具体操作：

- 输入：学生对某一推理步骤的回答（自由文本、表格或混合格式）
- 操作：LLM使用“extraction prompt”识别回答中的关键要素，并将其转换为JSON格式
- 输出：`{extracted_statements: [...], normalization_confidence: float}`

例子：在“假设生成 (DDx阶段)”中，学生可能写：

患者主要症状是胸痛和呼吸困难。我认为可能是心肌梗死、肺栓塞或气胸。
心肌梗死风险最高因为患者有高血压和吸烟史。

LLM的提取操作应输出：

```
{
  "extracted_ddx": [
    {"diagnosis": "心肌梗死", "rank": 1, "reasoning": "高血压和吸烟史"},
    {"diagnosis": "肺栓塞", "rank": 2, "reasoning": "呼吸困难"},
    {"diagnosis": "气胸", "rank": 3, "reasoning": "急性胸痛"}
  ],
  "extraction_confidence": 0.95
}
```

此操作的目的是将模糊的自然语言转化为清晰的"可被评价的陈述"，为后续的对比和打分做准备。

6.2 基准对比 (Benchmark Comparison)

目的：比较学生的陈述与"金标准答案"（参考答案），判断覆盖度、准确性、排序合理性。

具体操作：

- 输入：提取后的学生陈述 + 金标准答案（reference answer）
- 操作：LLM逐项对比，计算"匹配度"与"新颖性"
- 输出：`{match_scores: [...], novelty_flags: [...], ranking_alignment: float}`

例子：假设金标准答案是：

```
{  
    "correct_ddx": [  
        {"diagnosis": "心肌梗死", "rank": 1},  
        {"diagnosis": "肺栓塞", "rank": 2},  
        {"diagnosis": "食管痉挛", "rank": 3}  
    ],  
    "dangerous_exclusions": ["急性主动脉夹层"]  
}
```

LLM的对比操作输出：

```
{  
    "matches": [  
        {"student_diagnosis": "心肌梗死", "in_reference": true, "rank_alignment":  
        "exact"},  
        {"student_diagnosis": "肺栓塞", "in_reference": true, "rank_alignment":  
        "exact"},  
        {"student_diagnosis": "气胸", "in_reference": false, "rank_alignment": "N/A"}  
    ],  
    "missing_dangerous_diagnoses": ["急性主动脉夹层"],  
    "overall_coverage": 0.67,  
    "dangerous_omission_risk": "HIGH"  
}
```

这一步回答了**"学生的答案在多大程度上与专家共识相符"**。

6.3 多维度打分 (Multi-Dimensional Scoring)

目的：基于提取和对比的结果，对学生的表现在**过程维度**和**结果维度**分别给出数值分数。

对于**过程导向维度**，打分内容包括：

- **完整性 (Completeness)**：是否列出足够的诊断并涵盖不同类别

- **逻辑性 (Logical Coherence)** : 为每个诊断给出的理由是否合理、是否有跳跃
- **事实准确性 (Factuality)** : 提到的临床特征、风险因素等是否正确
- **表达简洁性 (Conciseness)** : 是否言简意赅

对于结果导向维度，打分内容包括：

- **准确性 (Accuracy)** : 排在首位的诊断是否正确；前三个中有没有正确答案
- **全面性 (Comprehensiveness)** : 是否遗漏高风险、高流行率的诊断
- **安全性 (Safety)** : 有没有明显危险或不合理的诊断
- **专业性 (Clinical Professionalism)** : 语言是否符合临床规范

例子：DDx阶段的打分：

```
{
  "process_oriented_scores": {
    "completeness": 3.5,
    "logical_coherence": 4.0,
    "factuality": 4.5,
    "conciseness": 3.0,
    "process_average": 3.75
  },
  "result_oriented_scores": {
    "accuracy_top1": 0.0,
    "accuracy_top3": 1.0,
    "comprehensiveness": 2.5,
    "dangerous_diagnosis_awareness": 0.0,
    "safety": 3.0,
    "result_average": 1.08
  },
  "overall_score": 2.41,
  "confidence": 0.92
}
```

6.4 理由与可解释性生成 (Rationale Generation & Explainability)

目的：为每个打分维度生成简明的自然语言“评分理由”，帮助学生和教师理解评分的依据。

具体操作：

- 输入：上述打分结果 + 学生陈述 + 评分维度定义
- 操作：LLM生成自然语言的评价，指出“这个维度表现好/不好的具体原因”
- 输出：`{dimension_rationales: {}, overall_feedback: str, improvement_suggestions: [str]}`

例子：

```
{
  "dimension_rationales": {
    "dangerous_diagnosis_coverage": "⚠️ 学生未提及'急性主动脉夹层'，这是高死亡率且症
```

状与心肌梗死相似的诊断。这是该回答中最严重的遗漏。",
"accuracy_top1": "学生排在首位的是心肌梗死，但根据病例全景（患者年轻、无冠心病风险因素），实际首诊应为肺栓塞。",
,
"overall_feedback": "你的诊断思路清晰，论证逻辑较好，但在诊断优先级与高危诊断的识别上有改进空间。",
"improvement_suggestions": [
"补充急性主动脉夹层作为候选诊断，并学习其快速识别特征",
"基于病例信息（年龄、风险因素）重新调整诊断优先级"
]
}

7. 具体案例：DDx阶段的完整工作流

病例背景：56岁女性，3小时急性胸痛伴呼吸困难，高血压、吸烟史

学生回答：

问题表征：56岁女性，急性胸痛伴呼吸困难，高血压吸烟史

初步诊断：

1. 心肌梗死
2. 肺栓塞
3. 气胸

我认为最可能是心肌梗死因为患者有高血压和吸烟。

LLM工作流输出：

步骤	输出
Step 1：提 取	结构化DDx列表，confidence 0.98
Step 2：对 比	与参考答案对标，发现排序错误、遗漏主动脉夹层
Step 3：打 分	过程分3.75，结果分1.08，整体2.41，confidence 0.92
Step 4：理 由	排序误：患者急性呼吸困难是PE的典型，不是MI；严重遗漏：未提主动脉夹层，高死亡率诊断； 建议：学习胸痛鉴别诊断流程图，特别是PE vs MI vs AD的快速区分

8. 置信度与多模型验证机制

仅靠单个LLM的评分是不够的。本工作流整合了近期文献中关于LLM评估可靠性的最佳实践。

8.1 多模型合伙 (Panel of LLM-Judges)

最近的OSCE LLM评分研究表明，当使用多个不同的LLM进行独立评分时，**模型间的一致性能显著提升评分的可信度**：单个模型与人类评分的 $\kappa=0.88$ ，但多模型ensemble（投票/平均）能将 κ 提升至0.95。[OSCE2024] 类似地，SCT Panel方法通过6个LLM模型的概率分布投票，为每个学生选项生成robust的评分。[SCT2025]

基于这些实证支持，在关键评估点，系统调用2–3个不同的LLM，生成多份独立的评分：

工作流：

```

学生回答
↓
LLM_A评分 → score_A, confidence_A
LLM_B评分 → score_B, confidence_B
LLM_C评分 → score_C, confidence_C
↓
计算模型间一致性 (agreement_score)
↓
IF agreement_score > 0.85:
    → 使用平均分，标记为"高置信度"，自动接受
ELSE IF agreement_score > 0.70:
    → 使用平均分，标记为"中置信度"，flag为"需教师审查"
ELSE:
    → 不自动评分，直接推送给教师进行人工评估

```

8.2 人工复核阈值与路由

近期工作强调，LLM评估系统中**缺少工程化的"何时信任模型、何时寻求人类介入"的决策规则**。[OSCE2024] 本工作流根据"置信度"与"一致性"设计了明确的路由策略：

```

IF confidence > 0.90 AND consistency > 0.80:
    → "完全自动评分，无需人工复核"
ELSE IF confidence > 0.80 AND consistency > 0.70:
    → "自动评分，但标记为'建议教师审查'"
ELSE IF confidence > 0.70 OR 涉及安全性相关维度 (如dangerous_diagnosis_awareness) :
    → "混合人机评估：LLM预评分，教师做最终决策"
ELSE:
    → "仅供参考，需完全人工评估"

```

这套策略确保了：1) 高置信的评分自动处理，提升平台效率；2) 低置信的评分有明确的人工介入路径；3) 涉及患者安全的维度（如危险诊断遗漏）永不自动决策。

小结：从原子操作到工程系统

第六到第八节设计的LLM评估工作流，将零散的LLM应用实验沉淀为一个标准化、可复用的评估系统：

- **四个原子化操作**（提取→对比→打分→理由）可独立使用，也可串联成完整流程；

- 多模型合议与置信度路由从文献中的"最佳实践"转化为工程规则；
- 人机协作框架确保了LLM的自动化优势与人类教学者的最终权力相平衡。

这套工作流既保证了评分的可靠性（通过多模型共识和一致性检查），也保证了透明性（通过结构化输出和自然语言理由）。更重要的是，它为教师和学生提供了可理解、可审查、可改进的评估过程。

第四部分：核心方向三 - 学生个性化评估

9. 从多维数据到个性化画像

9.1 原子数据来源与聚合策略

系统会从三个来源收集原子级评分信号，这些信号最终会被聚合为学生的个性化推理画像：

1) 传统量表维度

- CRI-HT-S (Focusing / Context Creation / Securing) ——来自病史采集任务
- Revised-IDEA (Interpretive Summary / Differential / Explanation / Alternatives) ——来自临床文书
- 各类OSCE、虚拟病人中的结构化评分维度

2) 智能系统的行为与诊断

- 交互型任务中的行为日志：问诊顺序、检查选择、假设变更轨迹等（在后续引入虚拟病人/问诊模块后启用）
- 错误模式：某学生在哪类病例上频繁遗漏哪类诊断

3) 本项目LLM-as-judge的输出

- 每次评分的多维分数（过程分、结果分、各维度得分）
- 附带的文字理由与问题识别
- 置信度标记（高/中/低置信度）

9.2 按推理步骤聚合

系统按六步框架做第一层聚合。例如对某学生的"DDx步骤"在10个病例上的表现：

```
DDx 步骤的聚合数据：  
├ 来自10个病例的LLM-as-judge评分  
│ ├ 过程分平均：3.2  
│ ├ 结果分平均：2.1  
│ └ 维度分解  
│   ├ completeness : 3.0  
│   ├ appropriateness : 3.5  
│   ├ dangerous_diagnosis_coverage : 1.5 ⚠  
│   └ confidence avg : 0.82  
└ 错误模式  
  ├ 遗漏\"主动脉夹层\"：4次  
  └ 诊断排序不当：5次  
└ 对比同侪  
  └ 你的avg process score : 3.2
```

```

  └班级avg : 3.5
  └你的dangerous_diagnosis_coverage : 1.5
    └班级avg : 2.8

```

9.3 纵向累积与趋势识别

系统在学期内多次收集这些聚合数据，形成时间序列，便于识别学习进度和平台期：

学生Alice的\"问题表征\"步骤纵向数据：

周次	过程分	结果分	语义精度	趋势
1	2.5	2.0	2.0	↗ 初始
2	2.8	2.3	2.2	↑
3	3.1	2.5	2.5	↑ 稳定改进
4	3.0	2.4	2.4	→ 平台期

10. LLM生成"临床推理画像"

10.1 画像输入与设计理念

LLM接收的输入是聚合数据 + 定义：学生各步骤的数值特征、班级对标、错误模式、纵向趋势，以及学生的基本背景（年级、专业等）。

在设计这一部分时，本项目并没有试图再引入一套全新的复杂理论，而是从教学和使用场景出发，选择了三个互补但直观的视角来组织画像输出：

- **Part A：推理风格**——回答“你是怎样思考的”：这一部分借鉴临床推理文献中的双过程模型，有比较明确的理论来源；
- **Part B：强项与短板地图**——回答“你相对擅长/薄弱在哪里”：更多是教学中常用的“长板/短板”视角，是一种直觉上合理、便于行动的总结方式；
- **Part C：案例驱动的具体观察**——回答“有哪些具体案例最能说明问题”：是面向教学实践的设计，突出少量关键、代表性的情境，帮助学生连接“抽象维度”和“真实病例”。

也就是说，**A**是“从理论到标签”，**B/C**更偏“从工程和教学常识到可用的结构化反馈”。三者合在一起，使得画像既有一定理论根基，又不过度复杂，便于学生和教师理解和使用。

10.2 画像输出：多层次叙述

在对学生推理风格进行定性时，本项目基于临床推理文献中的双过程诊断模型（Dual-Process Diagnostic Reasoning）。Yazdani等针对全科医学的批判性综述系统梳理了三类互补的认知机制：

1. 假设-演绎模型（Hypothetico-Deductive）：通过显式的线索-假设-验证循环进行诊断
2. 模式识别模型（Pattern Recognition）：基于快速、自动的模式匹配
3. 双过程模型：System 1（快速、自动、非分析型，对应模式识别）与System 2（缓慢、显式、分析型，对应假设-演绎）的交互

[Yazdani2017] 本节使用的"直觉型-分析型-混合型"分类，正是将这一双系统框架转译为教师和学生更易理解的教学标签——而非凭空发明新的理论体系。它将学生的推理风格映射为：

- **直觉型**：以System 1为主，快速模式匹配占主导，但在复杂/非典型病例上容易出错
- **分析型**：以System 2为主，显式假设验证占主导，但速度较慢、资源消耗大
- **混合型**：根据任务复杂度灵活切换两种系统，并有相互校正机制，是expert-like的状态

在这个理论框架之上，LLM生成的"临床推理画像"被组织为三个层次（Part A/B/C），分别对应**风格定性**、**量化短板**、**关键案例**三个互补的观察角度：

Part A：推理风格定性

基于学生在各步骤的特征，LLM将其分类为某种推理风格：

你的推理风格：混合偏分析型 (Analytical-Hybrid)

特征：

- 在信息采集阶段表现稳定（信息完整性：3.4/5），说明你倾向于系统性的收集而非随意提问。
- 在\"问题表征\"阶段，你常常试图包含过多细节，导致信息密度反而降低（vs同侪平均用词更简洁）。
- 假设生成时，排序通常合理，但危险诊断覆盖度明显低于同侪（你：1.5/5，班级平均：2.8/5）。
- 在\"假设评估\"阶段，你显示出较强的自我反思意识，这是你的强项，也是专家型推理的标志。

解释：你正在从\"快速直觉\"向\"系统分析\"转变。

Part B：强项与短板地图

在有了风格标签之后，仅告诉学生"你更偏分析型/直觉型"还不够具体，因此本项目增加了**第二个层次的画像**：围绕若干关键维度，给出学生当前的主要优势和优先改进点。

这里选择的维度并非来源于某一个特定量表，而是出于**教学和安全性角度的常识性考虑**：一方面，教师在指导学生时，往往会关心"这名学生在哪些环节已经做得不错，可以进一步强化"；另一方面，也会特别关注"哪些薄弱点如果不改，会直接影响患者安全或后续学习"。

因此，本项目在实践中采用了一个简单但清晰的结构：**列出 3 个最突出的强项 + 3 个优先级最高的短板**，并用已经在前文定义过的评分维度来支撑它们。这些维度本身来自第二部分的评估框架，但在这里不再做理论展开，只作为"阅读画像时的坐标系"。

三大强项：

1. 自我反思与校准能力
 - 在\"what doesn't fit\"识别上，班级排名前30%
 - 这表明你能够自我检查并调整假设

三个优先改进项：

1. ⚠ 危险诊断觉察度 (当前: 1.5/5, 班级平均: 2.8/5)
 - 具体表现: 4个胸痛病例中均未提及主动脉夹层
 - 风险等级: 高 (诊断遗漏直接威胁患者安全)
2. 问题表征的信息密度
 - Summary statements 平均92个词, 同侪平均58个词
 - 虽然详细, 但冗余信息占比35%
3. 诊断排序的一致性
 - 7个病例中, 你给出的DDx排序与专家参考不一致
 - 可能过度重视\"症状匹配度\"而忽视\"流行病学概率\"

换句话说, **Part B** 更像是一个"面向行动的汇总视图": 它不再讨论学生"是哪一类推理者", 而是告诉学生和教师:"如果这学期只有有限的时间, 你最值得花精力改进的是哪几件事"。

Part C : 案例驱动的具体观察

在风格标签和强/弱项摘要之后, 学生和教师通常还会问一个非常自然的问题:

"这些结论是从哪些具体表现推出来的?"

因此, 本项目增加了第三个层次: **用 2-3 个最有代表性的病例片段来"锚定"前面的抽象结论**。与 Part A 的理论来源不同, Part C 更多是基于教学经验的工程化选择:

- 不追求覆盖所有病例, 而是只选"最能说明问题"的少数案例;
- 每个案例都要清楚地回答三个问题: **发生了什么 → 这说明了什么问题 → 下一步可以怎么做**。

案例3 (56岁女性, 干咳2周, 在ARB治疗中):

- 你的答案: 列出了\"感染\"\"慢阻肺\", 但未提\"药物副作用\"
- 分析: ARB诱发咳嗽的可能性高于50%, 是首选诊断。
这提示你在\"用药史与副作用\"的关联上需加强学习。
→ 建议: 复习常见药物副作用速查表, 特别是ACE-I/ARB

案例7 (72岁男性, 胸痛伴血压升高):

- 你的答案: 正确识别了\"主动脉夹层\"并排在首位
- 积极反馈: 很好! 说明在这个病例上你成功应用了之前学到的知识。

从阅读体验上看, **Part C 是把 Part A 的风格标签和 Part B 的长短板"落地"到具体情境**: 学生可以看到, "原来我在危险诊断觉察度这项短板, 其实就体现在这些病例里的遗漏/识别差异上", 从而更容易接受和内化这些反馈。

11. 从画像到建议: 行动清单

11.1 短期建议 (Next Week)

⌚ 本周行动建议 (可在3-5小时内完成):

优先级 ① — 危险诊断快速掌握

- |- 任务：复习\"急性胸痛决策树\"中的Rule-Out顺序
- |- 资源：Chest Pain Rulout Protocol (资源库中附上)
- |- 验证方式：完成3个\"PE vs MI vs AD\"的对比推理练习
- |- 预期：这周结束前，胸痛类病例中\"主动脉夹层\"识别率达 >80%

优先级 ② — 问题表征精简

- |- 任务：用\"语义修饰符模板\"重写上周的3个summary statement
- |- 目标：将平均词数从92 → 65以内
- |- 验证：自评这三份改进版本

次要 ③ — 可选深化

- |- 病例重做：选择上周表现差的2个病例重新分析

11.2 中期建议 (This Semester)

█ 本学期持续跟踪与改进方向：

维度1：危险诊断觉察 → 个性化学习路径

- |- 周度任务：每周至少1个含\"致命但易遗漏\"诊断的病例
- |- 目标：到学期末，覆盖度从1.5 → 3.5以上
- |- 追踪指标：LLM每周自动计算覆盖度

维度2：诊断排序的流行病学调整

- |- 根本问题：过度依赖\"症状匹配\"，欠缺\"患者人口统计学\"
- |- 学习方式：每两周进行1次\"排序推理\"练习
- |- 工具：提供\"流行病学先验概率表\"供参考

维度3：强项深化 — 反思能力的进阶使用

- |- 挑选2-3个复杂病例做\"多诊断管理\"训练

11.3 未来可能的拓展方向

⌚ 超出本学期范围的潜在发展方向

- 专科分化分析：观察在\"内科病例\"vs\"外科病例\"中的推理风格是否差异
- 推理风格进化追踪：从当前的\"分析-混合型\"，观察是否逐渐演化为\"更expert-like的混合风格\"
- 团队协作推理：如果引入多人协同推理场景，分析你的\"假设提出\"vs\"假设验证\"在小组中的贡献

小结：从数据到决策

第九到第十一节完成了从多源评分数据到可读、可行动的学生报告的最后一公里。这一转化的核心价值在于：

- **不止于分数**：数值评分聚合为定性画像，便于学生理解“我是什么样的推理者”
- **不止于描述**：通过对比、错误模式和案例，诊断出学生具体学习问题
- **不止于反馈**：提供短期/中期/未来的分层建议和具体资源，让改进有方向和步骤

这样的设计既尊重了医学教学的复杂性，也提供了学生能够立即采取行动的明确方向。

第五部分：总结与讨论

12. 三个部分的闭环

临床推理教学框架 (Gap 1)

↓

定义了6步 + 2个评估维度

↓

LLM-as-Judge评估工作流 (Gap 2)

↓

在每一步每一维度上生成评分 + 理由

+ 多模型验证 + 置信度路由

↓

个人化评估画像与建议 (Gap 3)

↓

把多次评分聚合、生成学生的“推理轮廓”

+ 定性风格分类

+ 具体、分层的改进建议

↓

最终输出给学生与教师：

清晰、可读、可行动的个性化报告 + 学期进度追踪

13. 核心创新与差异所在

相对于现有平台的创新：

1. 统一的理论框架而非工具拼凑

- 从单一量表 (CRI-HT-S、Revised-IDEA) 进一步整合，形成从“线索获取”到“反思校准”的完整推理链
- 显式地融合了现有教学理论与LLM能力分析

2. 标准化的评估工程而非一次性实验

- 四个原子化操作 (提取、对比、打分、理由生成) 可在任意推理步骤和任意维度复用

- 多模型合议与置信度路由的工程规则被显式化，而非隐在黑盒中
- 从Hepius、Alteach、OSCE、SCT等先前工作中提炼出可复用的最佳实践

3. 从数据到决策的完整链路

- 不止于分数和图表，而是生成学生能理解、教师能行动的个性化报告
- 推理风格定性（基于双过程模型理论）、强项短板清晰映射、可执行建议与资源关联
- 这一点在现有平台中几乎完全缺失

平台演进路线：从基础评分到交互推理与长期追踪

为了避免在各处零散地讨论“哪一阶段做虚拟病人、哪一阶段启用哪些维度”，本节集中给出平台的演进路径，并说明每个阶段实际会用到框架中的哪些子集维度。

Phase 1：固定题目 + 结构化回答 + LLM 评分 + 多维报告 + 基础数据集

- **核心目标：**在不引入复杂交互的前提下，先把“六步框架 + 双维评估维度”在**固定病例 + 结构化回答**的场景里跑通，验证 LLM-as-judge 的可靠性，以及“多维评分 → 个性化画像 → 行动建议”这一整条闭环。
- **主要能力：**
 - 采用固定题目和结构化答题格式（如分步写出问题表征、DDx 列表、理由等）；
 - 在六步中的若干关键节点上调用 LLM 进行**过程+结果**双向评分，生成多维度分数与评分理由；
 - 汇总多次作答，产出**基础版“临床推理画像”**与可执行学习建议。
- **数据集形态：**
 - 约 200+ 个病例 × 多维评分标签，形成首个基础训练/评估数据集；
 - 尚不包含真实的问诊对话或复杂交互日志。
- **关于框架维度的说明：**
 - 第二部分中定义的一些评估维度（尤其是依赖真实对话过程的 CRI-HT-S 子维度，如“**聚焦性”“情境构建**”），在理论框架中已经给出，但在 Phase 1 只作为预留维度，不参与真实评分；
 - 个性化画像中的某些行为型特征（如问诊顺序、检查选择轨迹）同样在 Phase 1 不会出现，对应的数据与指标会在 Phase 3 才被真正启用。

Phase 2：UI 设计 + 数据集的分类整合（按专科/课程结构化）

- **核心目标：**在已有评分与报告能力的基础上，进行**产品化与课程适配**：让教师和学生在实际教学中更易使用，并让病例与指标在“课程/专科”维度上有清晰的结构。
- **主要能力：**
 - 设计并迭代学生/教师端界面：包括学生作答界面、结果可视化（雷达图、趋势图）、教师批量审查面板等；
 - 按系统/专科（如心内、呼吸、消化等）对病例和维度进行分类整合，支持按课程模块、教学单元布置作业；
 - 在保持 Phase 1 同步的前提下，逐步扩展基础病例数和学生使用规模。
- **数据集形态：**
 - 在原有基础数据上，增加“专科标签、难度标签、课程单元标签”等元数据；
 - 仍然以固定病例 + 结构化回答为主，尚未引入真实对话日志。
- **与框架的关系：**
 - 六步框架与双维评估维度的**定义保持不变**，只是通过 UI / 分类结构让这些维度更贴合课程实际；
 - 仍不启用依赖虚拟问诊过程的那部分子维度。

Phase 3：探索虚拟病人相关功能（问诊工作流 + 交互日志评估）

- **核心目标：**在前两阶段稳定运行的基础上，引入虚拟病人/问诊工作流，并真正启用那些依赖交互过程的数据与维度，验证“过程日志驱动的评估与反馈”的增量价值。
- **主要能力：**
 - 在现有平台上集成虚拟病人或半结构化问诊模块，让学生可以通过对话/操作来获取线索；
 - 记录问诊顺序、追问深度、检查选择与假设变更等交互日志；
 - 启用第二部分中理论上定义、但在 Phase 1-2 处于 dormant 状态的若干维度（如 CRI-HT-S 的“聚焦性”“情境构建”、与问诊策略相关的过程指标），并将其纳入 LLM-as-judge 的评分与画像生成。
- **数据集形态：**
 - 新增一部分“对话/交互型病例”数据：包括原始对话文本、系统事件日志、多维评分与人类标注；
 - 形成“固定题目 + 交互病例”并存的数据格局。
- **与前文的一致性说明：**
 - 报告中所有围绕虚拟病人和交互日志设计的评估标准，从一开始就是作为“框架中的一部分”提出的，但工程上计划在 Phase 3 才逐步上线；
 - 因此，在阅读第二、四部分时看到的某些“问诊/交互相关维度”，可以理解为在 **Phase 1-2 不参与运行、但在 Phase 3 有明确落地路径的“预留节点”。**

Phase 4：探索更多可拓展方向（深度集成与高级功能）

- **核心目标：**在核心功能稳定后，探索更长期、更系统的扩展方向，使平台从“课程内工具”演进为“跨学期、跨场景的推理学习基础设施”。
- **可能的拓展方向包括：**
 - 跨学期纵向追踪与预警：在多学期数据的基础上，分析学生推理风格的演化轨迹，建立“诊断困难预警”模型；
 - 团队协作推理：支持小组病例讨论场景，分析个体在团队中的角色贡献（假设提出 vs 假设验证等）；
 - **LMS 与教学系统集成**：与 Blackboard / Moodle / Canvas 等 LMS 对接，实现作业/成绩的自动同步；
 - **LLM 场景生成与个性化题库**：根据学生的弱项自动生成变体病例（例如专门强化“药物副作用识别”“危险诊断覆盖”等），形成高度个性化的训练路径。
- **与框架的关系：**
 - 不再扩展新的“维度种类”，而是围绕既有框架做**更长时间、更复杂场景**的应用与分析；
 - 虚拟病人、交互日志、固定题目三类数据在这一阶段被统一纳入一个长期、连贯的学习轨迹视角中。

14. 局限与未来方向

当前设计的局限：

1. 单个患者的推理而非复杂多诊断场景
 - 当前框架主要针对“单一初始主诉”的诊断推理
 - 对于多诊断患者或急性加重既往病的情况需要扩展
2. 定量数据的严谨性
 - 某些评估维度（如“语义精度”“信息密度”）仍需更多校准数据确保可靠性

- 多模型一致性的阈值（0.85/0.70等）是preliminary的，需通过pilot实验进一步调整
- 这些阈值的最终值应根据实际评分数据的分布来科学设定

3. 教师工作流与LMS集成（Phase 2及以后）

- 报告生成之外，还需设计教师如何快速批量处理、标记、反馈的界面
- 与学校现有LMS（Blackboard、Moodle等）的集成尚未在Phase 1范围内

4. 伦理与公平性考虑（需在系统设计文档中补充）

- LLM评分的潜在bias（如是否对非英语表述的学生不公平）
- 个人数据的安全性与隐私保护
- 算法透明度与学生对评分的申诉机制

未来可能的拓展方向：

1. 多诊断与管理决策推理

- 扩展框架以支持“患者既有肺炎又有免疫缺陷表现”等复杂场景
- 增加“管理方案选择”和“优先级排序”维度

2. 跨学期纵向追踪与预警

- 在数据积累充分的前提下，支持多学期的推理风格进化追踪
- 建立“诊断困难预警”模型，提前识别可能掉队的学生

3. 推理风格与临床工作流的关联

- 分析“直觉型推理”在急诊中的优势 vs 在复杂病例中的劣势
- 教导学生根据临床场景和时间约束灵活调整推理策略

4. LLM场景生成与个性化题库

- Phase 2后，利用LLM的强大生成能力，基于学生弱点动态生成针对性病例
- 例如：识别出学生在“药物副作用识别”上薄弱后，自动生成3个特别强调用药史的病例

第六部分：结论

临床推理教学正处于一个工具丰富但整合不足的时代。虚拟患者平台提供了完整的情景训练，ITS贡献了多维评分与学生建模，LLM研究展示了自动评分的潜力——但这些能力往往各自为政，难以形成一个以学生学习需求为中心的完整系统。

本项目通过三个环节的设计，试图在这个空隙中填补缺口：

第一步（第4-5节）是建立统一的理论基础。基于假说演绎模型、问题表征理论、临床推理文献中的双过程模型、以及现有量表和LLM研究的综合，我们提出“六步诊断推理 + 双维评估维度”的框架。这个框架既复用了CRI-HT-S、Revised-IDEA等已验证的工具，也显式地纳入了LLM在医学推理中的特有失败模式——例如对“dangerous diagnoses”的低覆盖、对流行病学的忽视等——使得理论基础更贴近实际教学需求。

第二步（第6-8节）是实现标准化、可复用的评分工作流。通过四个原子化操作（信息提取→对标对比→多维打分→理由生成），加上多模型合议与置信度路由，我们把零散的LLM评分实验沉淀成一个可工程化、可在不同

任务间复用的“评分中间件”。这套工作流既确保了评分的可靠性（通过多模型共识），也确保了评估的可解释性（通过自然语言理由），更重要的是，它为教师和平台开发者提供了一套可理解、可调整的评估设计语言。

第三步（第9-11节）是完成从数据到决策的最后一公里。这是本项目相对于现有ITS和评分工具的核心差异所在。我们不止于生成分数和图表，而是利用LLM强大的语言生成和综合能力，将多次、多维的评分数据聚合为一份可读、有洞察、可直接引导学生行动的个性化报告。报告包含三部分：推理风格的定性分类（基于临床推理文献的双过程模型）、学生的强项与短板清晰地图、以及短期/中期的具体可执行建议。这种设计既尊重了教学的复杂性（推理风格是多维的、动态的），也提供了操作性的明确方向（这周做什么、下周改什么、进度如何评估）。

本项目的意义在于：在一个已经有很多好工具的时代，通过系统的整合、标准化的工程设计、和面向学生学习需求的数据转化，让这些工具真正协同起来，形成一个连贯的、透明的、有反馈闭环的学习生态。这样的系统不需要从零开始重新发明轮子，而是在尊重现有研究成果的基础上，通过架构层面的创新，让它们真正为学生的学习成长服务。

参考文献

[Yazdani2017] Yazdani, S., Hosseinzadeh, M., & Hosseini, F. (2017). Models of clinical reasoning with a focus on general practice: A critical review. *Journal of Advances in Medical Education & Professionalism*, 5(4), 177–184. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5611427/>

[Gaber2025] Gaber, F., Shaik, M., Allega, F., et al. (2025). Enabling doctor-centric medical AI with LLMs through workflow-aligned tasks and benchmarks. *Nature Portfolio*. <https://www.nature.com/articles/s44401-025-00038-z>

[Awada2024] Awada, A., et al. (2024). An e-learning platform for clinical reasoning in cardiovascular diseases: a study reporting on learner and tutor satisfaction. *PMC*. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11385837/>

[BodyInteract2025] Virtual case reasoning and AI-assisted diagnostic instruction: an empirical study based on body interact and large language models. *BMC Medical Education*, 2025. <https://link.springer.com/article/10.1186/s12909-025-07872-7>

[Hepius2021] A Natural Language Processing-Based Virtual Patient Simulator for Training Clinical Diagnostic Reasoning. *JMIR Medical Education*, 2021. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8041050/>

[Alteach2022] Chen, M., & Li, Y. (2022). Intelligent virtual case learning system based on real medical records and natural language processing. *BMC Medical Informatics and Decision Making*, 22(1). <https://pubmed.ncbi.nlm.nih.gov/35246134/>

[VPDialogue2025] Using LLMs to Grade Clinical Reasoning in VP Dialogues. *ACL 2025 SIGDIAL*. <https://aclanthology.org/2025.sigdial-1.56.pdf>

[SCT2025] Teaching Clinical Reasoning in Health Care Professions Learners Using AI-Generated Script Concordance Tests. *JMIR Formative Research*, 2025. <https://formative.jmir.org/2025/1/e76618>

[OSCE2024] Large Language Models for Medical OSCE Assessment: A Novel Approach to Transcript Analysis. *arxiv*, 2024. <https://arxiv.org/abs/2410.12858>

[Elstein1978] Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). Medical Problem Solving: An Analysis of Clinical Reasoning. Cambridge, MA: Harvard University Press.

- [Regehr2018] Regehr, G., & Norman, G. (2018). Exercises in Clinical Reasoning: Take a Time-Out and Reflect. *Mayo Clinic Proceedings*, 93(5), 713-715. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5834975/>
- [DDxTutor2025] Lu, Y., et al. (2025). DDxTutor: Clinical Reasoning Tutoring System with Differential Diagnosis-Based Structured Reasoning. *ACL 2025 Main Conference*. <https://aclanthology.org/2025.acl-long.1495/>
- [Nori2025] Nori, H., et al. (2025). Towards accurate differential diagnosis with large language models. *Nature*. <https://www.nature.com/articles/s41586-025-08869-4>
- [Qiu2025] Qiu, S., et al. (2025). Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, 16. <https://www.nature.com/articles/s41467-025-64769-1>
- [Sim2025] Sim, S., et al. (2025). Critique of impure reason: Unveiling the reasoning behaviour of medical large language models. *eLife*. <https://elifesciences.org/articles/106187>
- [Zhang2025] Zhang, Y., et al. (2025). Automating Expert-Level Medical Reasoning Evaluation of Large Language Models. *Nature Digital Medicine*. <https://www.nature.com/articles/s41746-025-02208-7>
- [CLEVER2025] Liu, J., et al. (2025). Clinical Large Language Model Evaluation by Expert Review: Development and Validation of the CLEVER Rubric. *AI JMIR*. <https://ai.jmir.org/2025/1/e72153>
- [IDEA2014] The IDEA Assessment Tool: Assessing the Reporting, Diagnostic Reasoning, and Decision-Making Skills Demonstrated in Medical Students' Hospital Admission Notes. *Teaching and Learning in Medicine*, 26(3), 2014. <https://pubmed.ncbi.nlm.nih.gov/25893938/>
- [LLMEvalMed2025] Zhang, M., et al. (2025). LLMEval-Med: A Real-world Clinical Benchmark for Medical LLMs with Physician Validation. *ACM Findings (EMNLP 2025)*. <https://aclanthology.org/2025.findings-emnlp.263/>
- [Scoping2024] Large language models for disease diagnosis: a scoping review. *Frontiers in Digital Health*, 2024. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12216946/>