

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。本项目定位为一个**中间层平台**，通过三个核心创新来弥补现有的缺口：

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：

系统/研究	流程教学	LLM评分	个性化分析	主要贡献
Hepius	✓	X	~	早期ITS+学生模型
Alteach	✓	X	✓	多维评分与班级分析
虚拟患者平台	✓	~	~	真实情景仿真
OSCE LLM评分	~	✓	X	多模型共识机制
DDxTutor	~	✓	X	单任务LLM评分

可以看到，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、以学习评估为核核心目标”的统一平台。

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”。此外，这些工具的设计多基于医学教育和临床实践的传统经验，鲜少有人用LLM在诊断推理任务中的失败模式和能力分析来反思“学生应该特别在哪些步骤上获得训练与评估”。

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

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。

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：

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

定义与意义：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”。

理论基础来自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]

在教学中，应让学生显式陈述“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。

4.2 线索获取（Cue Acquisition）

定义与意义：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。

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]

教学上，这一步不应是“记忆”而是“策略”。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。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。

评估维度（来自CRI-HistoryTaking Scale）：

- **聚焦性（Focusing）**：问诊是否围绕主诉和危险信号
- **情境构建（Context Creation）**：是否主动追问时间轴、诱因、伴随症状、既往史
- **确认与总结（Securing）**：是否在结束前向患者进行总结并确认理解正确

4.3 问题表征（Problem Representation）

定义与意义：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”）。

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能更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。

评估要点：

- **Summary Statement质量**：是否抓住主诉、时间、关键症状和病程演变
- **语义精度**：是否用"急性/慢性"等医学术语准确描述
- **信息对齐**：是否遗漏了之前收集的关键信息

4.4 假设生成 (Hypothesis / Differential Diagnosis Generation)

定义与意义：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。

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]

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。

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

定义与意义：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。

原发性全科医疗诊断策略模型中的“细化阶段”描述了医生使用的多种策略：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]

关键维度：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。

4.6 反思与校准 (Reflection & Calibration)

定义与意义：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。

“Exercises in Clinical Reasoning”中强调“Time-out and Reflect”策略。eLife综述“Critique of impure reason”系统地区分了“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]

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]

具体评估维度：

- **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
- **事实正确性**：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
- **反思与自我校准能力**：是否能识别"what doesn't fit"并调整假设

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]

具体评估维度：

- **诊断准确性**：Top-1 accuracy和Top-N accuracy
- **诊断列表的合理与全面性**（Appropriateness & Comprehensiveness）
- **管理/建议的安全性**（Safety）
- **资源使用与流程效率**（Efficiency）
- **表达质量**（Clinical Communication Quality）

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：

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

LLM首先需要从学生的原始回答中，提取出可评估的"陈述"（statements），然后将其标准化为结构化形式。

具体操作：

- 输入：学生对某一步骤的回答（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，并将其转换为机器可读的格式
- 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"。

6.2 基准对比 (Benchmark Comparison)

一旦提取了学生的陈述，LLM需要将其与"金标准答案"进行对比，判断相似性、覆盖度、准确性。

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"
 - 匹配度 (Match Score)**：学生的答案在金标准中是否出现或等价出现
 - 新颖性 (Novelty)**：学生提出的诊断是否为金标准以外但仍合理的诊断
 - 排序合理性 (Ranking Coherence)**：学生给出的优先级排序是否与临床合理性一致
- 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"}
  ],
  "novelty_flags": [
    {"diagnosis": "食管痉挛", "flag": "novelty"}
  ]
}
```

```

"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)

基于提取和对比的结果，LLM对学生的表现在多个维度上给出分数。

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
- **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)

仅有分数是不够的。LLM需要为每个维度生成简明的"评分理由"。

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：

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 →

- 排序错：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：学习胸痛鉴别诊断的流程图，特别是PE vs MI vs AD的快速区分

8. 置信度与验证机制

仅靠单个LLM的评分是不够的。系统需要实现一套置信度判断与人机路由机制。

8.1 多模型合伙 (Panel of LLM-Judges)

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"
ELSE IF agreement_score > 0.70:
    → 使用平均分，标记为"中置信度"，附加人工复核推荐
ELSE:
    → 不自动评分，直接推送给教师评估
```

8.2 人工复核阈值与路由

系统根据"置信度"与"一致性"，自动决定是否推送给教师：

```
IF confidence > 0.90 AND consistency > 0.80:
    → "完全自动评分，不需人工复核"
ELSE IF confidence > 0.80 AND consistency > 0.70:
    → "自动评分，标记为'需要教师审查'"
ELSE IF confidence > 0.70 OR 存在安全性相关维度:
    → "混合人机评估：LLM预评，教师最终决策"
ELSE:
    → "仅供参考，需完全人工评估"
```

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：

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

- 虚拟病人系统中的交互日志：问诊顺序、检查选择、假设变更轨迹
- 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"：

```
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：

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

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。

10.2 画像输出：多层次叙述

在对学生推理风格进行定性时，本项目并非凭空发明标签，而是直接建立在临床推理文献中对分析型（hypothetico-deductive）、非分析型（pattern recognition），以及二者交互的双过程诊断模型（dual-process model）的区分之上。Yazdani等针对全科医学的批判性综述，系统梳理了假设-演绎模型、模式识别模型和双过程诊断模型作为三类互补的认知机制，明确区分了快速、自动、基于模式匹配的System 1（非分析型，对应pattern recognition/spot diagnosis）与缓慢、资源消耗大、基于显式假设检验的System 2（分析型，对应hypothetico-deductive/probabilistic reasoning）。该综述进一步总结了初始假设生成、模式触发（spot diagnosis）、受限排除（restricted rule-out）、逐步细化、概率推理等在初始、细化和最终诊断阶段的策略组合。[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——直觉型≈以System 1为主的pattern recognition/spot diagnosis占主导；分析型≈以System 2为主的hypothetico-deductive/rule-based占主导；混合型≈根据任务复杂度灵活切换、相互校正的expert-like状态——而非另起炉灶提出新的理论体系。

LLM的输出是结构化的自然语言报告：

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（CRI-HT-S Focusing: 3.4/5），说明你倾向于有目标的提问而非漫无目的。
- 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

聚焦前三个最突出的优点与三个需改进的方面：

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"

Part C：案例驱动的具体观察

从LLM评分理由中提取的关键观察——只列最有教学价值的2-3个：

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。

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

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

↓

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

+ 可执行建议

↓

最终输出给学生与教师：

清晰、可读、可行动的个性化报告

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
- 纳入LLM失败模式的分析维度

2. 标准化的评估工作流而非一次性实验

- 四个原子化操作 (提取、对比、打分、理由生成) 可在任意推理步骤复用
- 多模型合议与置信度路由的工程方案

3. 从数据到画像的最后一公里

- 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、强项短板清晰映射、可执行建议与资源关联

平台演进路线：

Phase 1 (当前midterm): 在文本化病例、结构化题库上验证核心框架

Phase 2: 整合虚拟病人交互，增加沟通能力评估

Phase 3: 按专科拆分数据集，计算专科特化分数

Phase 4: 跨学期纵向追踪与团队协作推理评估

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
- 多模型一致性阈值（0.85/0.70）等参数需通过pilot数据调整

3. 教师界面与工作流整合

- 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

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"维度

2. 跨学期纵向追踪

- 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
- 预测诊断困难的早期预警

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灵活调整推理策略

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、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（建立在临床

推理文献的双过程模型基础之上)、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. *ACL 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/>