下面先给出**统合后的"树状结构"**，再逐层说明每个节点是如何由哪些 paper 推导出来的（附标题+链接+简单解释）。

一、统合后的树状结构（教学步骤 × 评估维度）

> 目标：教学生完成**诊断性临床推理**，并对其过程与结果进行可操作的评估。

A. 临床推理流程（Process Tree）

1. 病例框架化与情境设定（Case Framing & Context)
   1.1 明确主诉与就诊目的
   1.2 把握就诊场景与不确定性水平（初级医疗 / 专科 / 急诊等）

2. 线索获取（Cue Acquisition）
   2.1 有焦点的问诊与体格检查
   2.2 结构化信息记录与重要信息确认

3. 问题表征（Problem Representation）
   3.1 提炼简洁的 Summary Statement
   3.2 使用语义修饰符 / illness scripts 组织线索

4. 假设生成（Hypothesis / DDx Generation）
   4.1 生成初步 DDx 列表（广度）
   4.2 合理的初步排序（根据流行病学+关键线索）

5. 假设评估与信息收缩（Hypothesis Evaluation & Narrowing）
   5.1 针对性补充检查/提问（诊断增益）
   5.2 使用概率/规则/关键鉴别点进行收缩
   5.3 形成诊断结论或"下一步管理决策"

6. 反思与校准（Reflection & Calibration）
   6.1 检查"不符合之处"（what doesn't fit）
   6.2 回顾推理路径与偏差（metacognition）
   6.3 外显化为书面记录/口头汇报

B. 评估维度（Evaluation Tree）

B1. 过程导向：推理迹象评估（Rationale-Based）

- B1.1 步骤覆盖度与完整性（per-step completeness）
- B1.2 线索–假设链接的合理性与一致性（logical coherence）
- B1.3 事实正确性（factuality）
- B1.4 表达的简洁/高效度（conciseness & efficiency）
- B1.5 反思与自我校准能力（reflective reasoning）

B2. 结果导向：结论评估（Conclusion-Based）

- B2.1 诊断准确性（Top-1 / Top-N accuracy）
- B2.2 诊断列表的合理与全面性（appropriateness & comprehensiveness）

- B2.3 管理/建议的安全性（safety）
- B2.4 资源使用与流程效率（efficiency / triage quality）
- B2.5 表达质量（clinical communication quality）

二、每一层如何由 paper 推导而来（+ 理论工具说明）

1. 病例框架化与情境设定

依据：

- *Models of clinical reasoning with a focus on general practice* – Yazdani 2017
  https://pmc.ncbi.nlm.nih.gov/articles/PMC5611427/
  这篇综述强调：
  - 全科与专科场景下，疾病谱、线索稀缺度、目标（是否必须下终诊）不同，会影响推理策略。
  - 例如在全科，医生往往只需要"是否需要转诊/复查"而非终极诊断。
    → 这说明在真正推理前，医生会**先对病例和场景做"框架化"**（这是所有模型的前提）。

- *Enabling doctor-centric medical AI with LLMs through workflow-aligned tasks and benchmarks* – Nature Digital Medicine
  https://www.nature.com/articles/s44401-025-00038-z
  文中将医生工作流分为"问诊前 triage 场景、门诊初诊、随访"等不同 task，LLM 评估基于这些上下文。
  → 支持在教学中显式区分"在哪个情境下做推理"。

理论工具（新增）："Case Framing"节点

- 作用：把散乱的病人输入放进"场景框架"中（初诊/复诊、急/慢等），影响后续线索选择与阈值。
- 教学意义：让学生先讲清楚"我现在在哪个临床场景、我要解决什么决策问题"，而不是直接列 DDx。

---

2. 线索获取（Cue Acquisition）

依据：

- Hypothetico-deductive model – Elstein 等经典研究，综述见：
  Models of clinical reasoning…
  https://pmc.ncbi.nlm.nih.gov/articles/PMC5611427/
  明确四步：cue acquisition → hypothesis generation → cue interpretation → hypothesis evaluation。
  → 线索获取是被多个模型视为首要环节。

- *An e-learning platform for clinical reasoning in cardiovascular diseases* – Awada
  https://pmc.ncbi.nlm.nih.gov/articles/PMC11385837/
  平台步骤：患者描述 → Anamnesis（病史采集）→ 体检 → 初步诊断。
  并用"must-ask questions"检验学生是否问到了关键线索。
  → 说明线索获取可具体拆成"聚焦问题 + 完整性检查"。

- *AIteach: Intelligent virtual case learning system based on real medical records*
  https://pubmed.ncbi.nlm.nih.gov/35246134/
  系统从真实病历中抽取关键问题/线索，评估学生的问诊、用语与信息覆盖程度。

理论工具（与 CRI/IDEA 对齐）：

- "Focusing / Completeness / Securing" 三子维度可以沿用你之前的拆分，本质都在这一层。

---

3. 问题表征（Problem Representation）

依据：

- *Exercises in Clinical Reasoning: Take a Time-Out and Reflect* – Mayo Clinic
  https://pmc.ncbi.nlm.nih.gov/articles/PMC5834975/
  明确指出：
  - 早期步骤是形成 **problem representation**：对关键症状、时间、严重程度的简洁抽象。
  - 使用 semantic qualifiers（如"acute vs chronic"、"progressive vs intermittent"）来抽象病情。
    → 支持把"线索解释/问题表示"作为独立步骤。

- *Teaching clinical reasoning: principles from the literature to help improve instruction from the classroom to the bedside*
  https://pmc.ncbi.nlm.nih.gov/articles/PMC11150937/
  总结了经典教学设计：
  - 先训练学生写 summary statement、列 problem list，再进入脚本匹配与 DDx。
    → 说明问题表征在教学中是可单独训练和评估的。

理论工具说明：Problem Representation / Semantic Qualifiers

- 作用：把长篇病史压缩成"高信息密度的一句话"。
- 教学上：这是区分"能不能像医生一样想"的关键标志，也是所有后续推理的基底。

---

4. 假设生成（Hypothesis / DDx Generation）

依据：

- Hypothetico-deductive model & pattern recognition model – 同上综述文
  https://pmc.ncbi.nlm.nih.gov/articles/PMC5611427/
  - 临床医生在少量线索后就会生成初步假设，再用新信息修正。
  - 生成多少、质量如何，会直接影响最终诊断。

- *DDxTutor: Clinical Reasoning Tutoring System with Differential Diagnosis-Based Structured Reasoning*
  https://aclanthology.org/2025.acl-long.1495/
  - 将每个病例拆成局部推理（线索→疾病支持度）与全局推理（DDx 排序）。
  - 对"DDx 的广度"和"排序合理性"都有显式评分。

- *Towards accurate differential diagnosis with large language models* – Nature
  https://www.nature.com/articles/s41586-025-08869-4
  - 直接评价 LLM 的 DDx Top-1 / Top-k 准确度以及列表合理性。

因此：
把"假设生成"细分为：

- 4.1 DDx breadth：是否涵盖所有关键候选诊断
- 4.2 DDx prioritisation：排序是否与病例特征、流行病学匹配

---

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

依据：

- *Models of diagnostic reasoning strategies in primary care* – 同一综述内的 model 6
  https://pmc.ncbi.nlm.nih.gov/articles/PMC5611427/
  - 描述在 refinement 阶段使用的多种策略：restricted rule-outs、stepwise refinement、probabilistic reasoning、clinical prediction rules 等。
    → 支持将"针对性信息收集 + 逐步收缩"作为独立步骤。
- *Teaching Clinical Reasoning in Health Care Professions Learners Using AI-Generated Script Concordance Tests* – JMIR Formative
  https://formative.jmir.org/2025/1/e76618
  - SCT 本质是：给出新线索，让学生判断每个诊断概率如何变化。
    → 体现的是"在不确定情境下，如何更新对各假设的信念"。
- *Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis* – NPJ Digital Medicine
  https://pmc.ncbi.nlm.nih.gov/articles/PMC12064692/
  - 在 triage、specialty referral、diagnosis 三个层面，看 LLM 能否在给定有限信息情况下做出合理决策。

因此：
这一块强调 3 个子点：

- 5.1 选择有"诊断增益"的新检查/问题（而非乱查）
- 5.2 用概率/规则/关键鉴别点更新假设
- 5.3 在不确定下给出"当前最合理的决策"（诊断 or 转诊 / 复查）

---

6. 反思与校准（Reflection & Calibration）

依据：

- *Exercises in Clinical Reasoning*
  https://pmc.ncbi.nlm.nih.gov/articles/PMC5834975/
  - 提出"Time-out and reflect"：刻意停下来问自己：

    > What else? What doesn't fit? Could there be more than one diagnosis?

  - 这是对整个推理过程的元认知检查。

- *Critique of impure reason: Unveiling the reasoning behaviour of medical large language models* – eLife https://elifesciences.org/articles/106187
    - 区分 reasoning outcome vs reasoning behaviour，并强调需要对模型推理过程进行诊断和修正。
- *Automating Expert-Level Medical Reasoning Evaluation of LLMs* https://www.nature.com/articles/s41746-025-02208-7
    - 通过 process supervision / reward models，奖励"中间步骤的反思与自我修正"。

理论工具：Metacognitive Reflection 节点

- 作用：让学生/模型不仅给出答案，还能评价自己的推理是否存在偏差或遗漏。
- 在教学与 LLM 评估中都被视为下一步关键能力。

三、评估维度：从工具与 LLM 研究中抽象

B1. 推理迹象评估（Rationale-Based）

主要依据：

- CLEVER Rubric – *Clinical Large Language Model Evaluation by Expert Review* https://ai.jmir.org/2025/1/e72153
    - 维度：Factuality, Clinical Relevance, Conciseness。
    - 专家按这些维度审查 LLM 输出的 reasoning / summary。
- *Critique of impure reason*（上）
    - 强调要分析"推理路径"的健全性，而不只是答案对不对。
- *Quantifying the reasoning abilities of LLMs on clinical cases* – Nature Communications https://www.nature.com/articles/s41467-025-64769-1
    - 显式统计每一个 reasoning step 的正确率、一致性，区分"过程正确但结论错" vs "过程就不对"。
- IDEA Assessment Tool https://pubmed.ncbi.nlm.nih.gov/25893938/
    - 评价 admission note 中的"Interpretive summary""Diagnostic reasoning""Decision-making rationale"等。
      → 实质就是对推理迹象打分。

归纳出的子维度：

- B1.1 per-step completeness：
    - 每个步骤（获取、表征、假设、评估）是否都有体现？
    - 对应 CRI / IDEA / SCT 中对"是否考虑了所有关键信息"的评分。
- B1.2 logical coherence：
    - 线索与假设之间是否有合理、非自相矛盾的链条？

    - 由 CLEVER 的“clinical relevance”+ Nature Comms 中对 step consistency 的分析推导。

- B1.3 factuality：
    - 每个推理步骤是否符合当前医学知识/指南。
    - CLEVER Factuality + Automating Expert-Level Reasoning 里的 process supervision 都在做这个。
- B1.4 conciseness & efficiency：
    - 输出是否信息密集、有信息增量，而不是反复啰嗦或无关。
    - CLEVER 的 Conciseness + 你原来的“Efficiency”维度。
- B1.5 reflective reasoning：
    - 有没有识别“what doesn’t fit”、自我校准。
    - 来自 Exercises in CR + Critique of impure reason + PRM 论文。

B2. 结论导向评估（Conclusion-Based）

主要依据：

- LLMEval-Med
  https://aclanthology.org/2025.findings-emnlp.263/
    - 采用“Usability Rate”：0–5 打分，≥4 视为 clinically usable。
    - 强调 correctness + usefulness + safety 综合。
- *Towards accurate differential diagnosis with large language models* – Nature
  https://www.nature.com/articles/s41586-025-08869-4
    - 直接看 Top-1 / Top-k 诊断准确率。
- *Evaluating LLM workflows in clinical decision support* – NPJ
  https://pmc.ncbi.nlm.nih.gov/articles/PMC12064692/
    - 评价 triage level（exact / within range）、specialty referral（top-3 是否包含正确）、diagnosis（at least one correct）。
- Disease diagnosis scoping review
  https://pmc.ncbi.nlm.nih.gov/articles/PMC12216946/
    - 总结现有研究的典型 outcome 指标：accuracy, safety, over/under-triage 等。

归纳出的子维度：

- B2.1 Diagnostic accuracy：Top-1 / Top-N
- B2.2 Appropriateness & comprehensiveness：
    - DDx 列表里每个诊断是否“合理地适合这个病例”；
    - 有无明显遗漏高概率疾病。
- B2.3 Safety：

    - 有没有危险建议、极端 under-triage。

- B2.4 Efficiency / triage quality：
    - 像 NPJ 论文那样，兼顾“准确 + 不过度消耗资源/检查”的评价。
- B2.5 Communication quality：
    - 在 AIteach / Hepius / e-learning 平台等系统中，常对学生的书写、汇报进行评分。
    - 可作为“结果”层面对临床沟通质量的一项附加指标。

四、最终可视化使用建议

你可以在 PPT 或一页纸中画一棵“倒三角树”：

- 顶层：**“Teach clinical diagnostic reasoning”**
- 左边大分支：**Process (步骤 1–6)**
- 右边大分支：**Evaluation (B1 过程 × B2 结果)**
- 每个节点旁边写上 1–2 个代表性 paper（标题+链接），你就可以一边展示结构，一边说“这个节点是从哪几篇 work 里抽象出来的”。

如果你愿意，我可以下一步直接给你一份“适合打印的一页树状图文本版本”（每个节点最多一行），方便你直接贴进 midterm 或做 slide。

1 2 3 4 5