arXiv:2603.01241v1 [cs.IR] 1 Mar 2026

TARSE：通过检索技能与经验进行测试时适应的推理智能体

Junda Wang [1] **Zonghai Yao** [1] **Hansi Zeng** [1] **Zhichao Yang** [1] **Hamed Zamani** [1] **Hong Yu** [2]

摘要

复杂的临床决策往往并非因为模型缺少事实而失败，而是因为模型无法可靠地在恰当的推理步骤中选择并应用正确的规程知识和先前示例。我们将临床问答视为一个智能体问题，其中包含两种明确且可检索的资源：技能，即可重复使用的临床操作规程，例如指南、规程和药理机制；以及经验，即来自以往已解决病例、经过验证的推理轨迹（例如思维链解答及其步骤级分解）。测试时，智能体从精心整理的库中检索相关技能和经验，并进行轻量级测试时适应，使语言模型的中间推理与临床有效逻辑对齐。具体而言，我们构建了：（i）一个技能库，其中的指南类文档被整理为可执行决策规则；（ii）一个经验库，其中收录了按步骤级转移编索引的临床推理链示例；以及（iii）一个感知步骤的检索器，用于为当前病例选取最有用的技能和经验条目。随后，我们利用检索到的条目对模型进行适应，以减少实例-步骤错位，并防止推理偏向缺乏依据的捷径。在医学问答基准测试上的实验表明，与强大的医学检索增强生成基线和仅使用提示的推理方法相比，我们的方法始终能够取得更好的结果。结果表明，明确区分并检索临床技能与经验，然后在测试时对模型进行对齐，是提升医学智能体可靠性的一种实用方法。 [1]

1. 引言

临床决策问题看似简单，却暗藏一条严格的分支规则（Jiang 等，2025年a；Sellergren 等，2025年；Xu 等，2025年）。例如，问题可能会询问：在一家小型医院，面对类似心脏病发作的表现，下一步该怎么做？此时虽然可以快速选择“溶栓”，但患者近期有脑出血史（Li 等，2023年；Wang 等，2025年b；2024年b）。采用分块式检索增强生成的大语言模型可以检索到若干真实片段（例如“迅速采取行动”和“出血风险”）（Jin 等，2025年；Zhao 等，2025年；Xiong等人，2024b；Zakka 等，2024年；Lozano 等，2023年；Wang 等，2024年a），但仍可能出错，因为它没有在恰当的步骤将硬性规则视为门控。相比之下，能够检索（i）明确编码分支规则（包括“禁止使用”条件）的程序性技能，以及（ii）在选择分支前检查该条件的类似经验证经验轨迹的医学领域智能体，就能够遵循正确的决策路径（如图1所示）。

这个例子展示了临床医生反复依赖的两种资源。首先，他们会运用可重复使用的程序性知识（例如指南、规程、决策规则和药理机制），这些知识可应用于许多病例。其次，他们会借鉴具体的既往经验（例如相似病例如何逐步得到解决，包括常见陷阱以及如何处理各种限制）。我们将这些资源形式化，以供医学智能体使用。在本研究中，我们将临床技能定义为可在不同患者间重复使用的结构化程序性知识。我们将临床经验定义为以示例思维链及其逐步分解形式表示的、来自既往已解决病例的经验证推理轨迹；这些表示揭示了每一步所需的推理或核查。将二者都视为明确的检索目标，使智能体能够在测试时调用它们，以指导多步推理。

尽管医学大语言模型和检索增强生成（Wu 等，2024a；Wang 等，2024a；Jiang 等，2025b）发展迅速，但当前系统很少支持这种技能与经验视角（Park，2025年）。首先，大多数医学检索增强生成流程检索的是无结构的文档片段，模糊了程序规则与特定病例推理轨迹之间的区别，使模型为了做出决策究竟应检索什么变得不明确（Xia 等，2025年；

[1]马萨诸塞大学阿默斯特分校，美国马萨诸塞州 [2]马萨诸塞大学洛厄尔分校，美国马萨诸塞州。联系邮箱：Junda Wang <jundawang@umass.edu>。

预印本。*2026*年*3*月*3*日。[1]代码将在录用后发布。

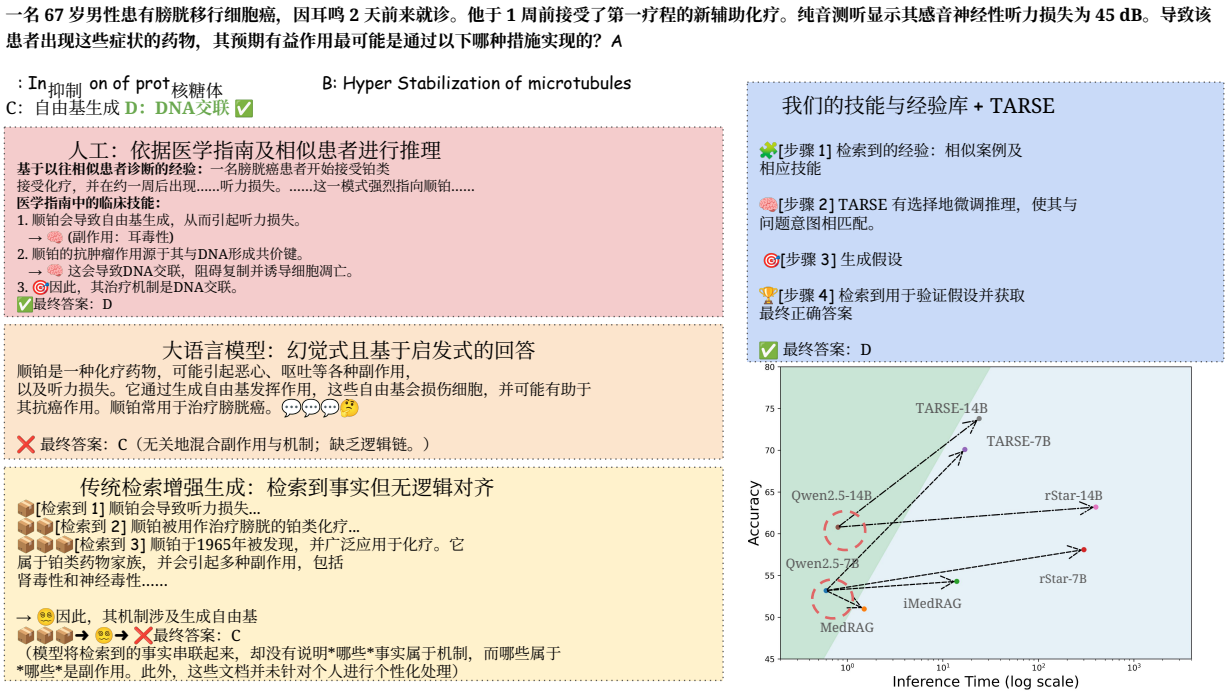


图*1.* **技能与经验库的动机与概述 + TARSE。（左）**临床问答示例，其中的关键难点是目标条件消歧：顺铂的耳毒性（自由基 → 听力损失）属于副作用链，而其抗肿瘤机制则属于治疗链（共价DNA结合 → DNA交联 → 细胞凋亡）。人类通过结合经验（回忆起一个相似的膀胱癌 +铂类化疗案例，该案例将症状与顺铂联系起来）和技能（可复用的指南/机制逻辑链），运用“机制与副作用”门控并选择正确分支（DNA交联，选项D），从而解决问题。相比之下，普通大语言模型往往遵循启发式关联，将毒性与机制混为一谈；传统的文本块检索增强生成则会检索到真实但缺乏结构的片段，这些片段无法与推理步骤对齐，因而导致同样的混淆。**（右）**我们的方法将这一流程具体化为：（1）检索相似的经验轨迹和相关技能；（2）进行轻量级测试时适应，使推理与问题意图对齐；（3）生成临时假设/逻辑链；（4）使用检索到的技能对其进行验证和修正，从而得出最终答案。右下角的图总结了准确率-延迟权衡，显示在推理时长相近的情况下，TARSE取得的提升大于仅使用思维链和传统检索增强生成基线方法。

Liang 等，2025年；Myers 等，2025年）。其次，即使检索到了相关信息，也常常与关键推理步骤不匹配：多跳临床问题通常取决于某一项缺失的检查（例如禁忌证、时间窗或排除标准），而全局检索无法可靠地在恰当的环节注入正确的规则。第三，主流医学检索增强生成方法（例如 MedRAG 风格的方法（Xiong 等，2024a）或迭代文本块检索（Xiong 等，2024b））并非为检索结构化形式的技能与经验而设计的（Ou 等，2025年）。指南是可执行的程序，而推理链是轨迹；将它们拆分成通用片段，可能破坏其决策结构，并促使模型对部分相关的片段进行“平均”，而不是遵循连贯的临床路径。这些局限促使我们设计一种测试时智能体，用于检索结构化的技能与经验，并逐步使模型的推理与之对齐。

我们提出一种检索增强医学智能体，明确分离并操作化临床技能与经验。我们从类指南文档构建技能库，将其组织为可复用的决策程序，并一个由步骤级转移索引的示例临床推理链经验库。在推理时，智能体对两个库执行步骤感知检索，选择最相关的技能条目和经验条目。随后，我们对检索到的条目进行轻量级测试时适应，以将语言模型的中间推理与临床有效逻辑对齐，从而减少实例-步骤错位，并防止推理偏向缺乏依据的捷径。我们在医学问答基准测试上评估该方法，并与强医学检索增强生成基线以及仅提示的推理方法进行比较。在各项任务中，我们观察到一致的提升，在需要显式检查和分支决策的多跳与约束密集型问题上增益更大。

贡献： i) 我们提出一种用于临床推理的技能—经验框架，并指出基于文本块的检索在结构上与这些资源不匹配；ii) 我们构建可重复使用的临床技能和示例经验库，并通过步骤级索引进行检索；iii) 我们提出一种测试时检索与适应框架，可提升医学领域

推理表现，并生成更具可解释性且以临床实践为依据的决策路径。

2. 相关工作

技能与经验在智能体文献中，技能是智能体可在不同任务或状态下调用的可复用流程。许多研究探讨如何通过分层控制、技能诱导和可调用例程库来发现并复用这些流程（Park，2025年；Wang 等，2025c）。技能之所以重要，是因为它们能将长期任务拆解为更短、更稳定的子程序，使智能体能够通过选择和组合已有知识实现泛化。另一类相关研究则在推理时将技能存储为可检索的材料，例如工具使用模板、程序草图、自然语言流程或检查清单（Jin 等，2025年；Xiong 等，2024b；Xia 等，2026）。这与具有书面规则的领域中的专家工作相契合。在医学领域，指南和规程发挥着这一作用，因为它们编码了旨在供不同患者重复使用的决策步骤和排除规则。已有系统会检索精选指导（Zakka 等，2024年），但大多数研究以提示为中心，将这些文档视为一般证据，而非用于指导逐步决策的结构化技能。

我们的方法遵循技能库视角，并将其具体应用于临床推理。我们把类似指南的流程表示为结构化技能条目（例如必需检查、分支条件、排除标准），并将其作为指定下一步的对象进行检索，而非松散的文本。我们还将技能与经验整合到一个测试时循环中。在强化学习中，经验回放和离策略方法会重用轨迹，以稳定学习并促进迁移（Zhang 等，2025年），近期研究也利用模型自行生成的轨迹及其结果作为监督（Wang 等，2025年a；Hu 等，2025年）。对于大语言模型，思维链和基于记忆的方法会重用推理轨迹来指导新问题的处理（Wei 等，2022年；Yan 等，2025年），但检索到的轨迹通常只是额外的语境，并未明确说明流程，因此模型可能检索到正确事实，却执行了错误步骤。

我们通过将经验 整理为经过验证的逐步推理轨迹，并按步骤级转移建立索引来解决这一问题，使检索针对缺失的步骤，而非仅关注整体相似度。在测试时，我们会同时检索 (i) 指南中相关的技能条目，以及 (ii) 匹配的轨迹片段；随后利用轨迹展示相似案例中的检查如何执行，并依据检索到的技能条目逐步引导模型。这针对的是基于分块的检索增强生成中一种常见的失败情形：检索到真实陈述，却遗漏必需检查，或在错误的环节应用硬性规则。

基于大语言模型与检索的推理大语言模型仍难以进行明确的多步推理以及检查并分支推理（Liu 等，2023年；Wu 等，2024b；Dziri 等，2023年）。提示方法（例如思维链、自洽性）可以提升流畅性（Wei 等，2022年；Wang 等，2022年），但在专家领域中，它们无法可靠地确保步骤顺序正确或决策可验证。检索增强生成通过注入外部证据来减少缺乏依据的主张（Lewis 等，2020年；Guu 等，2020年），然而，检索文本往往缺乏结构，或与需要支持的具体步骤不匹配（Ye 等，2023年；Xiong 等，2024a）。即使是医学检索增强生成变体（例如 MedRAG、iMedRAG），检索到的也大多是通用文档片段；一旦漏掉某项检查，就可能改变决策，因此这远远不够。我们的工作重新界定了临床推理的检索目标：智能体不再只检索文档片段，而是检索明确的程序性技能（可复用规则和检查清单）以及示例经验轨迹（经验证的逐步解决方案）。这使检索结果能够切实用于决策，并通过在需要的步骤提供恰当的规则或示例来减少步骤错配。

推理的测试时训练与适应 测试时训练/适应会在推理过程中利用测试时信号更新模型；这项技术在计算机视觉领域已得到充分研究（Sun 等，2020年； Behrouz 等，2024年），但在自然语言处理和大语言模型推理领域仍受到限制（Lee 等）。以往的目标通常是无监督的，与目标推理过程的对齐也较弱，因此由实例-步骤错位 导致的错误依然存在：模型在问题层面检索到相关证据，却仍在错误的步骤应用了错误的规则。我们通过将检索与针对智能体检索到的技能 和经验条目的轻量级测试时适应相结合来解决这一问题。检索到的材料提供了结构化的实例特定监督，使模型逐步推理与正确的决策逻辑对齐，而无需全局重新训练。

3. 问题表述

推理的隐链视角。 设 $q \in \mathcal{Q}$ 为一个问题， M_θ 为一个参数为 θ的小型语言模型。我们假设，回答问题的过程会依次经过一个未观测到的微观推理状态序列$\mathbf{z} = \langle z_1, \ldots\ldots, z_T \rangle \in \mathcal{Z}^T$（即潜在逻辑链），随后输出答案$y \in \mathcal{Y}$:

$$y \;=\; f_\theta(q, \mathbf{z}) \quad \text{with} \quad \mathbf{z} \sim p_\theta(\cdot \mid q).$$

每个相邻对 $\tau_t = (z_t \to z_{t+1})$ 都是一次W转移 （主张、变换或子推导）。错误通常出现在局部转移 $\tau_{t^\star}$，但由于$\mathbf{z}$ 未被外显，因此仍被隐藏。

朴素检索为何会错过出错步骤。标准检索以全局查询（即输入 q 或模型的草拟答案 y）为条件，从语料库 $\mathcal{C}$中得到文档集 $\mathcal{D}^{(q)} = \text{TopK}\big(r_\phi(q)\big)$ 和 $\mathcal{D}^{(y)} = \text{TopK}\big(r_\phi(y)\big)$

使用检索器 r_ϕ ： $\mathcal{X} \to 2^{\mathcal{C}}$。定义一个步骤对齐函数 A： $2^{\mathcal{C}} \times \mathcal{T} \to [0,1]$，用于衡量一组文档是否与特定转移 τ 直接相关或无关（例如，是否包含该步骤所需的前提、规则、实体关系或反证）。对于失败步骤 $t^\star$，实证结果表明，

$$A\big(\mathcal{D}^{(q)}, \tau_{t^\star}\big) \ll 1 \quad \text{and} \quad A\big(\mathcal{D}^{(y)}, \tau_{t^\star}\big) \ll 1,$$

因为与 q 或 y 的相关性是主题层面的，而非步骤索引层面的；因此，错误（或缺失）的微观前提会持续存在，并使错误不断传播。

目标。我们旨在找到一种方法：(i) 引出临时链 $\hat{\mathbf{z}} = \langle \hat{z}_1, \ldots, \hat{z}_T \rangle$，以反映模型针对该实例预期的中间状态；(ii) 按转移而非按 q 或 y进行检索，从而使证据在局部对齐：

$$\forall t \in \{1, \ldots, T-1\}: \quad \mathcal{E}_t = \mathrm{TopK}\big(r_\phi(q, \hat{\tau}_t)\big), \qquad \hat{\tau}_t = (\hat{z}_t \to \hat{z}_{t+1}).$$

以满足期望条件 $A(\mathcal{E}_t, \tau_t) \gg A(\mathcal{D}^{(q)}, \tau_t)$ $A(\mathcal{D}^{(y)}, \tau_t)$。利用步骤对齐证据 $\mathcal{E}_{1:T-1}$，模型将推理修订为更新后的链 $\mathbf{z}' = \langle z'_1, \ldots, z'_T \rangle$，并输出

$$y' = f_\theta\big(q, \mathbf{z}', \mathcal{E}_{1:T-1}\big),$$

以提升正确性和忠实性。

4. 经验与技能库构建

4.1. 技能库：指南式决策规则

我们首先通过从多种指南式来源（例如 CDC/NICE/WHO、教科书及其他医学领域语料库）中收集可复用的临床操作规程，初始化一个大型技能池。相关来源和语料库统计信息汇总于表10中。每个技能项 $s \in \mathcal{S}$ 都会被规范化为可执行规则格式（例如：条件 $\to$ 检验 $\to$发现 $\to$ 行动），以便在推理过程中作为步骤级“门控”进行检索和应用。

随后，我们从与问答配对的 LOGICALCHAIN 轨迹中构建经验库。每个条目都是一个三元组（q_i， y_i， $\mathbf{z}_i$），其中 q_i 是问题/病例， y_i 是已验证答案，而$\mathbf{z}_i$ 是按步骤索引的逻辑链。

人工评估。为验证所构建的经验条目，我们随机抽取 100 个提取出的实例，并请标注人员判断 LOGICALCHAIN 是否忠实且能够支持配对的问答。我们发现，抽取条目中有 88% 是正确的。

4.2. 经验库：逻辑链与问答轨迹

我们精心整理 LOGICALCHAIN 轨迹，使其满足三个原则：(i) 将推理分解为原子级、可验证的转换；(ii) 确保每个步骤都以支持性上下文中出现的实体为依据；(iii) 将推理链与问答紧密结合，使其具有实际作用，而非仅具装饰性。

设计原则1（结构化、按步骤索引的推理）。我们将一条经验轨迹表示为微观推理状态 $\mathbf{z} = \langle z_1 \to z_2 \to \cdots \to z_T \rangle$的序列，其中每个 z_t 都是一个微观推理状态，而每个转移 $\tau_t = (z_t \to z_{t+1})$则对应一个原子步骤。令$\mathcal{A}$ 表示一个最小原子转移类型 集合（例如，“前提→中间状态”、“中间状态→结论”），并令 $\mathcal{Z}$ 表示有效轨迹的集合。任何有效轨迹 $\mathbf{z} \in \mathcal{Z}$ 都可以分解为有限个原子转移$\mathbf{z} = \langle \tau_1, \tau_2, \ldots, \tau_{T-1} \rangle$ 的序列，其中 $\tau_t \in \mathcal{A}$。这种因式分解使经验能够（a）在转移层面上得到解释和验证，并且（b）按步骤进行寻址，从而支持步骤感知检索与对齐。

以医学领域中的一个具体例子来说，原子步骤可以是“症状 $\to$ 疑似药物类别”或“药物类别 $\to$治疗机制（规则应用）”。我们从自然包含已解决病例推理的来源中构建大型经验轨迹语料库（例如医学问答解释、教科书解答和其他经验证的推理依据）。如果只有事实语境（没有明确的推理轨迹），我们会在严格约束下合成按步骤索引的链（见附录A.3），并通过人工或自动方式筛选，以确保其忠实性。

由此得到经验库 $\mathcal{E} = \{(q_i, y_i, \mathbf{z}_i)\}_{i=1}^N$，其中每个$\mathbf{z}_i = \langle z_{i,1} \to \cdots \to z_{i,T_i} \rangle$都是一条经验证的推理轨迹，并与 q_i 和 y_i配对。

设计原则 2（基于实体的抽象）。设$\mathcal{U}$ 为临床实体的全集，并令 $\phi(c) \subseteq \mathcal{U}$从病例/语境中提取提及的实体 c。对于每个经验项（q_i， y_i， $\mathbf{z}_i$），我们强制进行实体对齐：

$$\mathrm{Ent}(\mathbf{z}_i) \subseteq \phi(c_i),$$

其中， c_i 是 q_i 的支持性上下文（例如解释性段落或病例描述）。这能确保每次转移都有依据，并减少无关步骤。

设计原则3（问答耦合轨迹）。我们要求答案必须由轨迹蕴含，即$\mathbf{z}_i \models y_i$。我们还会存储步骤级分解（例如将 τ_t 作为可检索的键），这样在测试时，模型就能根据缺失步骤 而非仅凭全局相似度检索经验。

5. TARSE

人类通过结合经验与技能来进行临床推理。在作答之前，临床医生会回忆一个相似的已解决病例（经验），以勾勒出一条暂定的推理轨迹（明确的逐步计划）。随后，他们会查阅

算法 1 TARSE：通过检索技能与经验进行测试时适应

输入： 临床病例 x，问题 q；基础大语言模型 M_θ；经验库 $\mathcal{E}$；技能库 $\mathcal{S}$；检索器 r_E, r_S；前m经验，前k项技能；测试时训练步骤 n，学习率 η；对齐阈值 τ。
输出： 最终答案 y；（可选）经验证的推理链 $\mathbf{z}^\star$；证据包 $\{\quad \widetilde{\mathcal{E}}, \widetilde{\mathcal{S}}\}$.
步骤 1：检索经验（相似的已解决病例） ▷ 图：步骤 1
$\mathcal{B} \leftarrow \mathrm{TopM}\big(r_E(q,\ x;\ \mathcal{E})\big)$ ▷ $\mathcal{B} = \{(q^{(i)},\ y^{(i)},\ \mathbf{z}^{(i)})\}_{i=1}^{m}$
步骤 2：TARSE =用于推理链对齐的轻量级测试时训练 ▷ 图：步骤 2
$\theta_q \leftarrow \theta$ ▷ 初始化查询自适应参数
对于 $j = 1, \ldots, n$ **执行**
  $\mathcal{L}_{\text{exp}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} -\text{日志}\ p_{\theta_q}\big(\mathbf{z}^{(i)} \mid q^{(i)},\ x^{(i)}\big)$ ▷ 对齐至检索到的经验轨迹
  $\theta_q \leftarrow \theta_q - \eta \nabla_{\theta_q} \mathcal{L}_{\text{exp}}$
结束循环
步骤 3：假设 / 临时链 + 步骤感知技能检索 ▷ 图：步骤 3
$\hat{\mathbf{z}} = \langle \hat{z}_1 \to \cdots \to \hat{z}_T \rangle \sim p_{\theta_q}(\cdot \mid q, x)$ ▷ 经验条件化假设生成
用于 $t = 1, \ldots, T-1$ **执行**
  $\hat{\tau}_t \leftarrow (\hat{z}_t \to \hat{z}_{t+1})$ ▷ 当前转移
  $\mathcal{S}_t \leftarrow TopK\big(r_S(q,\ x,\ \hat{\tau}_t;\ \mathcal{S})\big)$ ▷ 检索候选技能
  $\widetilde{\mathcal{S}}_t \leftarrow \{s \in \mathcal{S}_t :\ \text{对齐}(s;\ q,\ x,\ \hat{\tau}_t) \geq \tau\}$ ▷ 仅保留与步骤对齐的指南/方案规则
循环结束
步骤 4：验证后作答 */ ▷ 图：步骤 4
$y \sim p_{\theta_q}\ (\cdot \mid q, x, \mathbf{z}\ \{\widetilde{\mathcal{S}_t}\}_{t=1}^{T-1})$ ▷ 最终答案遵循经验证的治疗链
返回 y **(和** $\mathbf{z}^\star = \mathbf{z}$**)**。

指南式程序（技能），以验证其中脆弱的步骤（例如禁忌证门控、作用机制与副作用的区分，或所需的时间窗口），从而通过与步骤对齐的验证来纠正局部错误。整个过程如图2所示

我们通过 TARSE 将这一工作流程付诸实践。**阶段 A（经验对齐）**。给定一个测试查询，我们从经验库（QA +经验证的推理轨迹）中检索一小批相关的已解决病例，并进行轻量级测试时适应，使模型在检索到的经验分布下，针对当前实例生成一条临时的、特定于该实例的推理链。**阶段 B（技能验证）**。随后，我们将临时链中的每个转移作为步骤查询，以检索技能（指南/方案规则），这些规则可直接应用于相应步骤；再将规则反馈给模型，以验证或修正对应的转移。这一两阶段流程与人工实践相仿：首先以既往经验为计划奠定基础；然后，在单个端到端推理循环中运用技能验证程序正确性。

阶段 A：测试时经验对齐与临时链。为了显式呈现模型针对当前查询的隐藏推理，我们进行简短的、以邻居为条件的经验对齐，使模型在检索任何技能之前，先确定一条明确的、按步骤索引的临时链。给定一个测试查询 q，链感知检索器 r_ϕ（参数为 ϕ）会返回一个由链三元组 $\mathcal{S}(q) = \{(q^{(i)},\ y^{(i)},\ \mathbf{z}^{(i)})\}_{i=1}^{m}$ 构成的小型示例集，大小为 m，其中每个 $\mathbf{z}^{(i)} = \langle z_1^{(i)} \to \cdots \to z_{T_i}^{(i)} \rangle$ 都是一条与问题 $q^{(i)}$ 和答案 $y^{(i)}$配对的按步骤索引的逻辑链。从基础模型参数 θ 出发，我们以学习率 η_{cot} 执行若干梯度步，仅 将输出分布 $\pi_\theta(\cdot)$ 对齐到显式链：

$$\theta_q \ = \ \theta \ - \ \eta_{\text{cot}} \nabla_\theta \left(\frac{1}{m} \sum_{i=1}^{m} -\log \pi_\theta\big(\mathbf{z}^{(i)} \mid q^{(i)}\big) \right),$$

从而得到查询自适应参数 θ_q。随后，我们为该实例生成一条临时链：

$$\hat{\mathbf{z}} = \langle \hat{z}_1 \to \cdots \to \hat{z}_T \rangle \sim \pi_{\theta_q}(\cdot \mid q),$$
$$\hat{\tau}_t = (\hat{z}_t \to \hat{z}_{t+1}), \quad t = 1, \ldots, T-1,$$

其中，T是生成链的长度，$\hat{\tau}_t$ 表示其第t个转移。通过将分散的潜在假设归并为具体序列 $\hat{\mathbf{z}}$，阶段 A 使该实例的中间状态和转移变得可观察，并提供阶段 B 使用的步骤查询 $\{\hat{\tau}_t\}$。

阶段 B：步骤感知技能检索与验证。 为纠正步骤局部错误，我们根据临时链中每个转移检索技能 （指南/协议规则），而不是根据原始问题检索通用文档片段。给定所 elicited 的链 $\hat{\mathbf{z}}$ 及其转移 $\hat{\tau}_t$ for t=1, ..., $T-1$，技能检索器 r_S 根据（q, $\hat{\tau}_t$）构造步骤条件化查询，并返回一个较小的候选技能项排序集合：

$$\mathcal{S}_t \ = \ \mathrm{TopK}_k\big(r_S(q, \hat{\tau}_t)\big) \quad \text{with} \quad \mathcal{S}_{1:T} = \{\mathcal{S}_t\}_{t=1}^{T-1}.$$

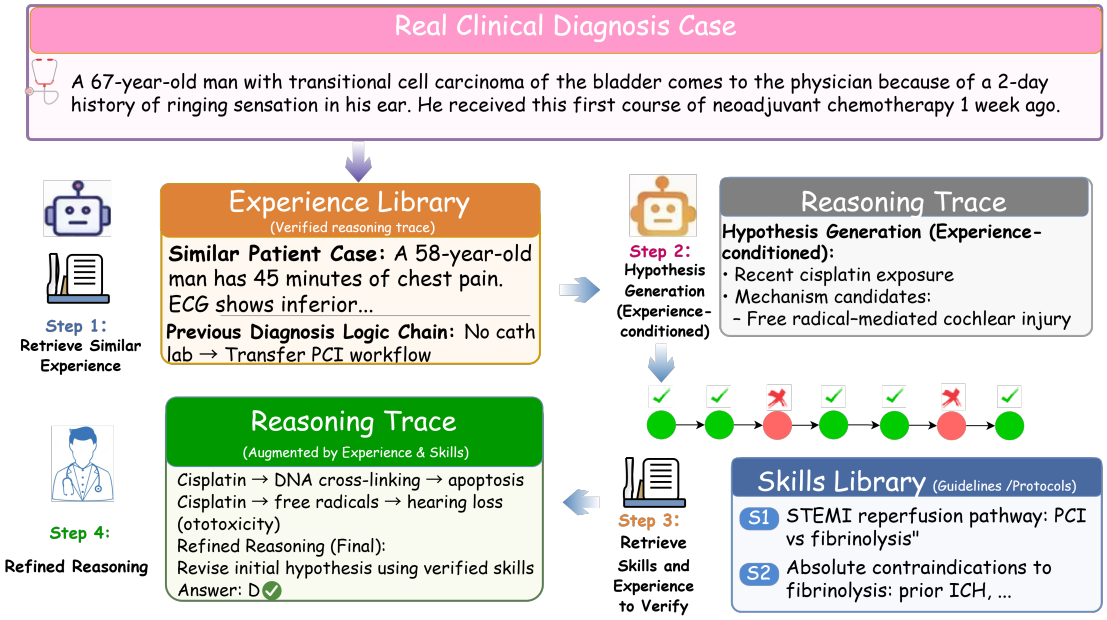


图2. **TARSE过程**。 从真实的临床诊断问题（顶部）出发，智能体首先从经验库中检索一条相似的经验证的经验轨迹（步骤 1），然后执行经验条件化假设生成以提出候选机制（步骤 2）。接下来，它从技能库中检索技能（例如，指南/方案规则），并连同支持性经验条目一起，用于验证关键转移并执行“门控”检查（步骤 3），从而生成一条精炼的推理轨迹，清晰地区分副作用链（顺铂 → 自由基 → 耳毒性/听力损失）与治疗机制链（顺铂 → DNA交联 → 细胞凋亡）。最终答案遵循经验证的治疗链（步骤 4）。

可选地，步骤级对齐评分器 $s_\psi(s;\ q,\ \hat{}\tau_t) \in [0,1]$ 会根据阈值 τ 对每个候选池进行筛选：

$$\widetilde{\mathcal{S}}_t \ = \ \{\, s \in \mathcal{S}_t \ : \ s_\psi(s\,;\, q, \hat{\tau}_t) \geq \tau \,\}, \qquad \widetilde{\mathcal{S}}_{1:T} = \{\widetilde{\mathcal{S}}_t\}_{t=1}^{T-1}.$$

然后，我们将模型条件化于（i）阶段 A 中所用的检索到的经验小批次以及（ii）步骤对齐技能包，并生成最终答案（可选地输出一条经验证的/更新后的链）：

$$y' \ \sim \ \pi_{\theta_q}\big(\ \cdot\ \big|\, q,\, \hat{\mathbf{z}},\, \widetilde{\mathcal{S}}_{1:T},\, \mathcal{B}\big).$$

通过根据每次转移 $\hat{}\tau_t$ 进行检索，同时将技能作为规则来检索，阶段 B 提供了验证或修复链条中脆弱环节所需的精确门控/标准/机制，使模型的逐步计划与临床有效流程保持一致。

6. 实验

6.1. 数据集与模型

数据集我们在三个医学问答数据集上进行评估：MedQA（Jin 等，2021年）（美国执业医师资格考试步骤 1/2CK/3，2022年6月–2023年3月）、MedMCQA（Pal 等，2022年）（涵盖 21 个学科的 19.4 万道选择题）以及 MMLU–Medical（Hendrycks 等，2020年）（九个医学领域）；此外，还在两个通用多跳基准测试上进行评估：Multi-HopQA（Song 等，2018年），这些任务需要跨文档证据和分步验证。

模型 我们的骨干模型是 Qwen2.5（7B/14B）（Hui 等，2024年）。基线方法包括指令微调大语言模型（Qwen2.5-Instruct）、检索增强生成系统（MedRAG、i-MedRAG）以及测试时推理方法（rStar–Qwen2.5）；为进行更广泛的比较，我们还纳入了开源推理模型和专有大语言模型。在医学领域和通用多跳任务中，TTT–RAG 始终优于开源模型和检索增强生成基线方法。

6.2. 评估设置

为展示检索恰当经验并结合相应技能的重要性，我们设计了实验，直接将我们的方法与传统推理和标准医学检索增强生成基线进行比较。

推理轨迹质量（经验对齐）。我们评估检索经验并与之对齐能否提升模型生成的推理轨迹质量。具体而言，我们提示模型“先思考，再作答”，并将不进行自适应的思维链基线和强大的开放模型（Qwen2.5 7B/14B/32B、Llama3.1 8B、Mistral-7B、Phi-4 14B）与我们的经验对齐版本（TARSE-14B）进行比较。我们同时开展自动评估和人工评估。自动评估方面，我们使用字符串相似度和步骤级蕴含（原子覆盖率）衡量生成轨迹与标准参考答案之间的对齐程度。人工评估方面，我们从不同模型变体生成的轨迹中随机抽取100条，请标注人员对其以下方面进行评分：（1）逻辑连贯性，以及（2）

表*1.* MedQA 上的组件使用情况、准确率和延迟。时间 / *Q* 报告了相对于 Qwen-2.5（CoT）基线增加的每题延迟。

	医学领域			通用领域
模型	MedQA	MedMCQA	MMLU	MultiHopQA
Qwen2.5-7B-Instruct	53.2	54.1	70.3	26.3
Qwen2.5-14B-Instruct	60.8	55.7	75.2	32.5
MedRAG-Qwen2.5-7B	51.0	56.9	68.1	33.2
i-MedRAG-Qwen2.5-7B	54.3	57.6	74.4	35.6
rStar-Qwen2.5-7B	58.1	55.9	69.8	29.1
rStar-Qwen2.5-14B	63.2	56.7	77.2	34.1
TARSE-7B	70.1	62.1	75.4	45.1
TARSE-14B	**73.8**	**64.3**	**78.5**	**48.5**

表*2*。 模型在医学领域 与通用 领域的表现对比。

方法	使用思维链	使用检索增强生成	使用测试时训练	准确率（MedQA）	时间/题（秒）	词元
Qwen-2.5（思维链）	✓	✗	✗	55.1	+0s	+0
Qwen-2.5-rStar	✓	✗	✗	58.1	+497s	+491,132
仅测试时训练	✗	✗	✓	59.1	+24s	+37,826
仅检索增强生成技能（MedRAG）	✗	✓	✗	58.7	+0s	+2,641
思维链+检索增强生成技能（iMedRAG）	✓	✓	✗	62.5	+26s	+23,143
测试时训练+经验	✓	✗	✓	59.3	+25s	+39,177
测试时训练+经验+检索增强生成技能	✓	✓	✓	70.1	+32s	+56，214

领域正确性。我们报告相对于基线方法的胜率。

答案准确率（基于技能的验证与纠正）。我们评估在恰当步骤检索技能能否纠正错误并提高最终准确率。我们采用两种互补的设置。跨领域比较：在医学领域数据集（MedQA、MedMCQA、MMLU–Medical）和通用多跳数据集（MultiHopQA、2Wiki）上，我们使用相同的骨干模型（Qwen2.5–7B/14B），并比较：（i）思维链基线方法（Qwen2.5 CoT、rStar）；（ii）问题级RAG（MedRAG）；（iii）交互式思维链+RAG（i–MedRAG）；以及（iv）我们的方法：先检索经验以形成暂定推理链，再检索步骤对齐技能，在作答前对其进行验证和修复。

受控消融实验： 我们进一步通过三种配置隔离各个组件，并确保骨干模型、提示和解码方式一致：**设置 1** 测试时训练（无经验监督、无技能检索；直接给出答案）、**设置 2** 经验对齐的TTT（逻辑）（采用经验/轨迹监督；仍直接给出答案），以及**设置 3**TARSE（经验对齐 +步骤感知技能检索与验证）。如果提升主要出现在设置 3 中，且相较于（i）–（iii）在不同领域均能保持，我们便将改进归因于步骤对齐技能验证，而非更长的推理过程或问题级检索。

效率指标。 除准确率外，我们还从以下两方面报告效率：（i）每题增加的实际耗时，以及（ii）每题增加的词元数，其中

表 *3.* 不同模型在MedQA数据集上的推理质量表现。

模型	ROUGE-1	原子覆盖率	人工
Qwen2.5 7B	0.2841 ± 0.0063	0.3541 ± 0.0063	31%
Llama3.1 8B	0.2640 ± 0.0146	0.3215 ± 0.0270	39%
Mistral-7B	0.2617 ± 0.0012	0.3211 ± 0.0021	30%
Phi-4 (14B)	0.2438 ± 0.0012	0.3213 ± 0.0040	45%
Qwen2.5 14B	0.2429 ± 0.0646	0.5437 ± 0.0682	41%
Qwen2.5 32B	0.2225 ± 0.0003	0.5936 ± 0.0017	41%
TARSE-14B	**0.3318** ± 0.0130	**0.6112** ± 0.0210	–

token开销是相对于CoT基线测量的。

7. 结果

7.1. 主要结果

推理轨迹质量。 我们评估检索经验（来自已解决病例的已验证推理轨迹）并与之对齐，能否提升模型针对新问题生成的推理轨迹质量。与普通CoT基线相比，轻量级测试时适应能带来一定的稳定性提升，但仍逊于经验对齐变体；MedQA 上的轨迹质量比较展示了这一点（表 3）。在强大的非测试时训练开源模型中，ROUGE-1 集中在一个较窄的区间内（Qwen2.5-7B0.2841 ± 0.0063、Llama3.1-8B 0.2640 ± 0.0146、Mistral- 7B 0.2617 ± 0.0012、Phi-4-14B 0.2438 ± 0.0012），而步骤级蕴含（原子覆盖率）在相同骨干模型上主要随规模扩大而提升（Qwen2.5-14B $0.5437 \pm 0.0682 \rightarrow$ Qwen2.5-32B 0.5936 ± 0.0017），这表明更大的模型

可能会断言更多有依据的事实，却未必能生成步骤对齐的推理轨迹。相比之下，我们的经验对齐流水线（TARSE-14B）取得了最出色的自动推理轨迹质量，在 ROUGE-1 0.3318 ± 0.0130 和 原子覆盖率0.6112 ± 0.0210方面均表现最佳。这表明，检索相似已解决病例中的经验并据此进行适应，有助于模型更忠实地呈现步骤索引推理轨迹。

答案准确率。表2 按医学领域和通用领域划分，总结了各项性能。在不同数据集上，我们的方法优于强基线方法，这表明，将基于经验的推理与步骤对齐的技能验证相结合，对多跳和约束密集型问题是有效的。

跨领域规律。一个一致的趋势是，在正确性取决于在恰当时机执行特定程序性检查（例如排除标准、禁忌证门槛，或机制与副作用的区分），而非单纯回忆事实的数据集上，TARSE带来的提升更大。这解释了为何TARSE不仅在医学领域（MedQA/MedMCQA/MMLU）表现更好，也能迁移至通用多跳基准测试（MultiHopQA）：瓶颈往往在于步骤选择和验证，而这与领域无关。

为何问题级RAG并不足够。将MedRAG和i-MedRAG与TARSE进行比较表明，如果检索证据未与失败转移对齐，检索更多文本并不能稳定地带来更高的准确率。在表1中，仅RAG（MedRAG）相较于思维链基线略有提升，但仍远低于测试时训练+思维链+RAG。这一差距支持了我们的核心主张：主要的失败模式并非缺少证据本身，而是证据错置——在问题层面检索到的相关片段往往过于粗略，无法纠正某个特定的错误步骤。

为什么 TARSE 能胜过更长的“思考”。大规模测试时扩展（例如 rStar）会增加推理时的探索，但不会改变模型所视为的决策瓶颈。当错误源于混淆两条相近的因果链（治疗机制与不良反应机制），或跳过门槛条件时，采样更多链条可能会放大同一种错误启发式。相比之下，TARSE 明确地（i）将临时链锚定到相似的经验证经验上，然后（ii）注入与步骤对齐的技能作为可执行规则，以约束脆弱环节；这是一种性质迥异的纠正机制。

实践启示。 总体而言，结果表明一条简单的部署指南：如果某个基准测试（或真实临床环境）中，有许多问题的标签会因是否进行某项中间检查而发生变化，那么加入步骤感知技能验证 比增加思维链长度或检索更多非结构化段落更可靠。

7.2. 消融研究。

表 1 分别考察了思维链、检索增强生成和测试时适应（TTT）的贡献。在 TTT 期间加入步骤索引监督后，模型在不使用检索作答时取得了幅度不大但稳定的提升（仅 TTT：59.1，相比 Qwen-2.5（思维链）：55.1；以及 TTT+思维链：59.3），这表明与经验对齐有助于模型生成更清晰的中间前提。

主要的提升出现在通过检索使用技能验证推理链之后：TTT+CoT+RAG 在 MedQA 上达到 70.1，优于 CoT+RAG（iMe- dRAG：62.5）。这支持了我们的假设：准确率提升的关键驱动因素是步骤对齐的技能验证，而非单纯延长思维链。

7.3. 效率

表1揭示了准确率—延迟—词元权衡。纯思维链（55.1， +0秒）表现最差。不借助检索、仅通过 rStar 增加“思考”量，提升也很有限（58.1，+497秒），表明效率较低。单独使用任一组件都能带来适度提升：仅使用检索增强生成，准确率提升至 58.7，且不增加耗时（+3.6， +0秒）；仅使用测试时训练则达到 59.1，但会带来适度开销（+4.0， +24秒）。将思维链与通用交互式检索增强生成（iMedRAG）结合，可带来更大幅度的提升（62.5，+26秒），表明推理与文档访问之间具有协同作用。

除了延迟之外，表1 还报告了相对于思维链每个问题增加的词元成本。总体而言，完整的测试时训练+思维链+检索增强生成仍然是最佳权衡（70.1， +32秒）：与重度测试时扩展基线方法相比，它提升了准确率，同时将延迟和 token开销都保持在适中水平。

8. 结论

我们提出 LOGICALCHAIN + TARSE，这是一种检索增强推理框架，通过测试时适应整合结构化逻辑链，从而提升大语言模型的性能。TARSE 通过动态对齐模型的推理过程与检索到的推理路径，显著提升了模型在医学和数学基准测试上的表现，超越了现有的检索增强基线方法和指令微调基线。实验结果表明，在推理过程中引入结构化推理路径，能够同时提高模型输出的准确率和逻辑连贯性。这些所见证明，逻辑引导的实例特定适应能够有效提升大语言模型处理复杂任务的推理能力。

影响声明

本研究探讨如何通过在测试时检索两类经过筛选的资源，提升推理智能体的可靠性：（1）临床技能，即源自指南和规程、可重复使用的逐步规则；（2）临床经验，即来自已解决病例、经过验证的推理轨迹。智能体会针对当前推理步骤检索合适的技能和经验，再运用轻量级测试时适应，使其中间推理与这些资源保持一致，并避免采用缺乏依据的捷径。这有助于改善医学问答中的决策支持，尤其适用于需要明确检查和分支判断的多步骤问题。该方法也可以推广到其他流程明确且积累了大量过往已解决案例轨迹的领域，例如安全检查清单、调试、合规审查和科学工作流程。与此同时，医学领域的部署需要谨慎。错误或过时的技能、有偏差的经验轨迹或错误的检索结果，仍可能导致错误答案。过度依赖检索到的工具输出也可能误导用户。这些系统应作为决策支持工具，而非自主诊断工具；并且应通过临床安全规程进行测试，为检索到的条目保留审计日志，同时实施严格的隐私控制措施。

参考文献

Behrouz, A.、Zhong, P.与Mirrokni, V. Titans：在测试时学习记忆。*arXiv*预印本 *arXiv:2501.00663*，2024年。

Cobbe, K.、Kosaraju, V.、Bavarian, M.、Chen, M.、Jun, H.、Kaiser, L.、Plappert, M.、Tworek, J.、Hilton, J.、Nakano, R.等。训练验证器解决数学应用题，2021年。网址 *https://arxiv. org/abs/2110.14168*，9，2021年。

Dziri, N.、Lu, X.、Sclar, M.、Li, X. L.、Jiang, L.、Lin, B. Y.、Welleck, S.、West, P.、Bhagavatula, C.、Le Bras, R.等。信念与命运：Transformer在组合性上的局限。神经信息处理系统进展，36：70293–70332，2023年。

Guu, K., Lee, K., Tung, Z., Pasupat, P.和Chang, M. 检索增强语言模型预训练。载于国际机器学习会议，第3929–3938页。PMLR，2020年。

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., 等。Olympiadbench：一项具有挑战性的基准测试，旨在通过奥林匹克级双语多模态科学问题推动通用人工智能的发展。*arXiv* 预印本 *arXiv:2402.14008*，2024年。Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., 和 Steinhardt, J. 衡量大规模多任务语言理解。*arXiv* 预印本 *arXiv:2009.03300*，2020年。

Hu, M., Zhao, P., Xu, C., Sun, Q., Lou, J.-G., Lin, Q., Luo, P., 和 Rajmohan, S. Agentgen：通过环境和任务生成增强基于大语言模型的智能体的规划能力。发表于第*31*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集第*1*卷，第496–507页，2025年。Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., 等。Qwen2. 5-coder 技术报告。*arXiv* 预印本 *arXiv:2409.12186*，2024年。Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J. T., 等。Hulu-med：一种旨在实现整体医学视觉语言理解的透明通才模型。*arXiv* 预印本 *arXiv:2510.08668*，2025年a。

Jiang, X., Fang, Y., Qiu, R., Zhang, H., Xu, Y., Chen, H., Zhang, W., Zhang, R., Fang, Y., Ma, X.等。Tc–rag：图灵完备检索增强生成在医学大语言模型系统上的案例研究。载于第*63*届计算语言学协会年会论文集（第*1*卷：长文），第11400–11426页，2025年b。

Jin, B., Zeng, H., Yue, Z., Yoon, J., Arik, S., Wang, D., Zamani, H.和Han, J. Search-r1：利用强化学习训练大语言模型进行推理并借助搜索引擎。*arXiv*预印本*arXiv:2503.09516*，2025年。

Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H. 和 Szolovits, P. 这位患者患有什么疾病？一个基于医学考试的大规模开放域问答数据集。*Applied Sciences*，11(14):6421，2021年。

Lee, Y., Kim, D., Kang, J., Bang, J., Song, H. 和 Lee, J.- G. Ra-tta：面向视觉语言模型的检索增强测试时自适应。载于第十三届国际学习表征会议。

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T. 等。面向知识密集型自然语言处理任务的检索增强生成。神经信息处理系统进展，33:9459–9474，2020年。

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. Chatdoctor：一种利用医学领域知识在大语言模型 meta-ai（llama）上微调的医学领域聊天模型。*Cureus*，15(6)，2023年。

Liang, S., Zhang, L., Zhu, H., Wang, W., He, Y., 和 Zhou, D. RGAR：用于事实感知医学问答的递归生成增强检索。*arXiv* 预印本 *arXiv:2502.13361*，2025年。Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., 和 Cobbe, K. 让我们逐步验证。载于第十二届

国际学习表征会议，2023年。

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q. 和 Zhang, Y. 评估 ChatGPT 和 GPT-4 的逻辑推理能力。*arXiv* 预印本 *arXiv:2304.03439*，2023年。

Lozano, A., Fleming, S. L., Chiang, C.-C., 和 Shah, N. Clinfo. ai：一个开源的检索增强大语言模型系统，利用科学文献回答医学问题。载于 *2024*年太平洋生物计算研讨会，第 8–23 页。World Scientific，2023年。Myers, S., Dligach, D., Miller, T. A., Barr, S., Gao, Y., Churpek, M., Mayampurath, A., 和 Afshar, M. 评估检索增强生成与长上下文输入在电子健康记录临床推理中的表现。 *arXiv*预印本 *arXiv:2508.14817*，2025年。Ou, J., Huang, T., Zhao, Y., Yu, Z., Lu, P., 和 Ying, R. 利用电子健康记录进行经验检索增强，实现准确的出院问答。 *arXiv*预印本 *arXiv:2503.17933*，2025年。Pal, A., Umapathi, L. K., 和 Sankarasubbu, M. MedMCQA：用于医学领域问答的大规模多学科选择题数据集。载于健康、推理与学习会议，第 248–260 页。PMLR，2022年。Park, S. Instruct-SkillMix：用于大语言模型指令微调的强大流程。硕士论文，普林斯顿大学，2025年。Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., 等。MedGemma技术报告。 *arXiv*预印本 *arXiv:2507.05201*，2025年。Song, L., Zhang, Y., Wang, Z., 和 Gildea, D. 一种用于AMR到文本生成的图到序列模型。 *arXiv*预印本 *arXiv:1805.02473*，2018年。Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., 和 Hardt, M. 通过自监督进行测试时训练，以应对分布偏移下的泛化问题。载于国际机器学习会议，第 9229–9248 页。PMLR，2020年。Team, A.-M. AI-MO AMC验证数据集。

https://huggingface.co/datasets/AI-MO/aimo-validation-amc，2024年。访问日期：2024年5月9日。Wang, H., Wang, J., Leong, C. T., 和 Li, W. STECA：用于大语言模型智能体学习的逐步轨迹校准。 *arXiv*预印本 *arXiv:2502.14276*，2025年a。Wang, J., Yang, Z., Yao, Z., 和 Yu, H. JMLR：联合医学大语言模型与检索训练，以增强推理以及专业问答能力。*arXiv*预印本 *arXiv:2402.17887*，2024年a。

Wang, J., Yao, Z., Yang, Z., Zhou, H., Li, R., Wang, X., Xu, Y. 和 Yu, H. Notechat：一个基于临床笔记生成的合成患者—医师对话数据集。载于计算语言学协会论文集：*ACL 2024*，第 15183–15201 页，2024b。

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A. 和 Zhou, D. 自洽性提升了语言模型中的思维链推理能力。*arXiv* 预印本 *arXiv:2203.11171*，2022年。

Wang, X., Li, J., Chen, S., Zhu, Y., Wu, X., Zhang, Z., Xu, X., Chen, J., Fu, J., Wan, X., 等。华佗-26m：大规模中文医学问答数据集。载于计算语言学协会人类语言技术年会：*NAACL 2025* 所见，第 3828–3848 页，2025b。Wang, Z. Z., Gandhi, A., Neubig, G. 和 Fried, D. 为智能体任务诱导程序化技能。*arXiv* 预印本 *arXiv:2504.06821*，2025c。Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., 等。思维链提示能激发大语言模型的推理能力。载于神经信息处理系统进展，35:24824–24837，2022年。Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F. 和 Grau, V. 医学图检索增强生成：通过图检索增强生成实现安全的医学大语言模型。*arXiv* 预印本 *arXiv:2408.04187*，2024a。Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J. 和 Kim, Y. 推理还是复述？通过反事实任务探索语言模型的能力与局限。载于*2024*年北美计算语言学协会分会年会：人类语言技术论文集（第 *1* 卷：长篇论文），第 1819–1862 页，2024b。Xia, P., Chen, J., Wang, H., Liu, J., Zeng, K., Wang, Y., Han, S., Zhou, Y., Zhao, X., Chen, H., 等。SkillRL：通过递归技能增强强化学习进化智能体。*arXiv* 预印本 *arXiv:2602.08234*，2026。Xia, S., Lin, H., Wu, Y., Zhou, Y., Li, Z., Wan, Z., Xing, X., Zheng, Y., Li, X., Shan, C., 等。MedRek：利用键感知提示对医学大语言模型进行基于检索的编辑。*arXiv* 预印本 *arXiv:2510.13500*，2025。Xiong, G., Jin, Q., Lu, Z. 和 Zhang, A. 医学领域检索增强生成基准测试。载于计算语言学协会 *ACL 2024* 所见，第 6233–6251 页，2024a。

Xiong, G., Jin, Q., Wang, X., Zhang, M., Lu, Z. 和 Zhang, A. 通过迭代式后续问题改进医学领域的检索增强生成。载于生物计算 *2025*：太平洋研讨会论文集，第 199–214 页。World Scientific，2024b。

Xu, W., Chan, H. P.,Li, L., Aljunied, M., Yuan,R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z. 等。灵枢：一种用于统一的多模态医学理解与推理的通用基础模型。*arXiv* 预印本 *arXiv:2506.07044*，2025年。

Yan, S., Yang, X., Huang, Z., Nie, E., Ding, Z., Li, Z., Ma, X., Kersting, K., Pan, J. Z., Schütze, H., 等。Memory-r1：通过强化学习增强大语言模型智能体管理和利用记忆的能力。*arXiv* 预印本 *arXiv:2508.19828*，2025年。Ye, R., Zhang, C., Wang, R., Xu, S., 和 Zhang, Y. 自然语言就是图所需的一切。*arXiv* 预印本 *arXiv:2308.07134*，2023年。Zakka, C., Shad, R., Chaurasia, A., Dalal, A. R., Kim, J. L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., 等。Almanac——面向临床医学的检索增强语言模型。*NEJM AI*，1(2):AIoa2300068，2024年。Zhang, K., Chen, X., Liu, B., Xue, T., Liao, Z., Liu, Z., Wang, X., Ning, Y., Chen, Z., Fu, X., 等。通过早期经验进行智能体学习。*arXiv* 预印本 *arXiv:2510.08558*，2025年。Zhao, X., Liu, S., Yang, S.-Y., 和 Miao, C. MedRAG：通过知识图谱引导的推理增强医疗保健副驾驶的检索增强生成。发表于*2025*年世界互联网大会 *ACM* 论文集，第4442–4457页，2025年。

A. 附录

A.1. 鲁棒性：数学领域评估。

对于数学推理，我们在四个广泛使用的基准测试上进行了全面评估：MATH（Lightman 等人，2023年）、AMC 2023（Team，2024年）、Olympiad Bench（He 等人，2024年）和 GSM8K（Cobbe 等人，2021）。我们还进行了进一步的消融研究，以更好地了解 TARSE 框架的行为：

测试时训练期间的小批次大小的影响。我们系统地调整 TTT 中使用的适应批次大小（例如，1、2、4、8 个样例），以评估检索到的监督信号数量如何影响适应质量和最终准确率。这项消融研究有助于揭示适应速度与泛化性能之间的权衡。

并行多查询测试时适应。为模拟多用户推理场景并评估可扩展性，我们将并发用户查询分组（例如，3、6、9、12 个查询），并使用拼接检索批次进行联合适应。这种并行策略旨在摊销 TTT 的成本，同时实现更高效的资源共享。我们测量这种方法对不同并行度下准确率和适应稳定性的影响。

并行化实验。为探究逻辑链整合和测试时并行化的不同配置如何影响数学推理性能，我们开展了两组消融研究，结果见表4。

在图3中，我们改变**保留比例**，即测试时训练期间检索到的逻辑链——样本数从20到100，并观察到准确率单调上升。这表明，在适应过程中保持更高比例的逻辑推理步骤，能显著提升模型解决复杂符号问题的能力，使准确率从60.9%提高到63.2%。我们还评估了并行用户批次大小的影响：将来自3、6、9或12个用户的检索批次聚合，并共同用于测试时训练。我们发现，适度的并行度（6个用户）通过降低推理路径的方差来提升性能，而过高的并行度（12个用户）会引入噪声并降低准确率。这表明并行效率与推理保真度之间存在权衡，并凸显了在符号推理任务中进行受控批次级适应的重要性。

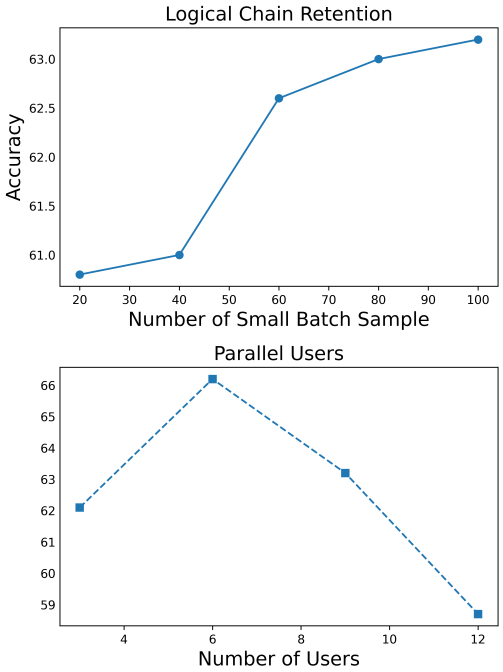


图 *3*。 TARSE 的消融研究。（上）逻辑链监督中所使用的保留小批次样本数量的影响。（下）批次式测试时训练期间并行用户数量变化的影响。

表 4。模 各数学基准上的el性能。rStar-Math在所有指标上均表现出强劲性能 s.

模型	MATH	AMC2023	Olympiad	GSM8K
Qwen2-7B-Instruct	49.6	25.0	0.04	82.3
NuminaMath-7B-COT	55.8	27.5	0.03	76.3
rStar-Qwen2	60.4	30.0	20.0	**87.2**
TARSEs	**63.2**	**30.0**	**24.7**	87.0

A.2. 错误分析

为了更好地了解不同检索策略的局限性，我们对基线检索增强生成方法与提出的 TARSE 框架进行了错误分析。我们使用 Qwen2.5-14B 作为基础模型，检索文档并直接生成答案，并将其与 TARSE 的测试时适应检索和推理进行比较。我们随机抽取了 100 个测试案例，并将每个错误预测人工归入以下四种错误类型中的一种或多种：(1) 无关或缺失证据，(2) 误导性证据，(3) 推理能力不足，以及 (4) 遗忘知识。

ColBERT-RAG 设置中，有 15 个错误是由无关或缺失文档导致的，6 个源于误导性内容，20 个归因于推理失败，另有 6 个涉及遗忘的背景知识。相比之下，TARSE 将这些数字分别降至 7、0、12 和 5。这些结果表明，TARSE 不仅提高了检索相关性，还减少了与推理和知识相关的错误，说明测试时适应能为下游推理提供更可靠且与语境更对齐的输入。

A.3. 数据构建流程

我们将文档分为两类：（1）**由专家撰写且包含明确推理轨迹的文档**，例如临床指南、医学资格考试解析或教科书习题解答。这些文档通常包含结构化逻辑，可提取或稍作改写，以形成可用的推理链。（2）**不包含明确推理的事实性文档**，包括教科书段落、精选问答数据集和维基百科等百科资源。这些文本提供了丰富的医学知识，但需要进一步处理，才能揭示其底层推理结构。详情见表5

对于第（2）类，我们通过使用结构化指令模板提示 GPT-4 来合成逻辑链。每个输入都包含一段医学事实段落，并要求模型生成一条逐步展开的逻辑链，将语境与结论联系起来，同时避免较大的逻辑跳跃。我们设置以下约束，以确保：

- 逻辑链以文档中的关键事实或条件开头；
- 包含中间推理步骤（例如生理机制、诊断步骤或治疗依据）；
- 最后一步在逻辑上推导出与医学领域相关的结论（例如诊断、机制、治疗）；
- 语言保持通用且简洁，除非临床上有必要，否则避免不必要的具体信息（例如年龄、实验室检查数值）。

一个具有代表性的提示如下：

> “以下文本包含一条医学事实。你的任务是生成一条逻辑链，解释其背后的推理过程。请简洁但完整。返回格式：逻辑链：A →B →C”

这一过程使我们能够将静态医学知识转化为动态推理轨迹，以便在推理时检索并对齐。

表5. 逻辑规则生成的提示模板。

系统提示	以下文本包含医学信息。你的任务是根据 所呈现的医学知识制定规则，确保每条规则都能全面体现 证据文本中的信息，不遗漏任何重要细节。
1. 明确区分条件与结果	每条规则都应清楚说明哪些项目属于条件、患者特征 或前提条件（例如特定诊断、临床发现或具体情形） ，哪些属于结果、措施或建议。始终将条件列在 规则的开头，后面是建议的措施或结果。
2. 针对复杂证据	如果一条证据陈述包含多个不同方面，请为 每个部分创建单独规则。这样可以确保每个概念都得到完整且清晰的体现，避免 无关想法被合并到同一条规则中。根据需要重复使用证据，以便 每条不同的规则.
3. 并列项与逻辑结构	在每条规则中按逻辑顺序排列各项。 • 对于依据现实逻辑具有自然顺序的项目（例如患者 特征 →诊断检查 → 治疗建议），请遵循这一顺序。 • 如果各项并列且没有严格顺序，则将它们标示在同一层级 在规则内，而不强加顺序（例如，“PET获益 → 诊断、分期、 治疗反应”）。
4. 现实逻辑顺序	在适用时，应依据现实世界中的关系遵循自然顺序，而不是 而非句子顺序。推荐的结构如下: • 条件或要求 • 诊断检查 • 发现或结果 • 治疗建议或措施 • 随访或监测计划（如适用） 无论这些要素在 原始文本。
5. 重构指南	确保生成的证据和规则集合能够全面涵盖指南， 以便根据这些部分准确地重构指南。尽可能使用 原始指南文本中的精确措辞。
6. 涵盖每一项关键信息细节	应体现证据中的每个重要术语、短语或概念。每条规则 在完整体现所有关键细节的同时，应尽可能简洁。 冗余，但要包含所有必要信息。
7. 直白表达避免使用条件句	使用清晰、直接的语言，避免过度使用“如果”或“当.....时”等条件句。 每条规则都应逻辑完整且自成一体，传达完整的语境 信息 .

避免

”

8. 输出格式
医学文本:
{medical_text}
示例规则格式：

需包含的关键词: {key_words}

正确：
证据: “一名患者出现腹痛，随后接受了CT扫描，检查显示有脓肿。治疗包括抗生素。”
规则: 腹痛 → CT扫描 → 检测到脓肿 → 治疗：抗生素。

错误：
证据: “一名患者出现腹痛，随后接受了CT扫描，检查显示有脓肿。治疗包括抗生素。”
规则: 腹痛 → 检测到脓肿 → CT扫描 → 治疗：抗生素。

返回格式

- 证据:
- 规则:
- 证据:
- 规则:

仅返回指定的输出格式，不要添加评论或解释。

A.4. 检索器训练与评估

A.4.1. 逻辑链检索器训练

我们使用两阶段检索-重排序流程检索逻辑链（）。首先，使用 ColBERT 为每个查询 q检索排名前 K 的候选逻辑链。然后，使用交叉编码器对这些候选项重新排序。该流程兼顾了检索效率与重排序效果，利用 ColBERT 快速进行稠密检索，并借助交叉编码器更强的语义匹配能力。具体而言，交叉编码器经过训练，可将查询 q （通常是用户问题）映射到对应的逻辑链 l^+，这些内容均从共享文档 d中采样。每个训练样本都构造成一个三元组（q，l^+，$\{l_i^-\}_{i=1}^N$），其中 l^+ 是黄金推理路径，l_i^- 是从同一文档 d内相邻逻辑链中采样的困难负样本。训练目标是最小化以下对比损失:

$$\mathcal{L} = -\log \frac{\exp(\text{Sim}(q, l^+))}{\exp(\text{Sim}(q, l^+)) + \sum_{i=1}^{N} \exp(\text{Sim}(q, l_i^-))}, \tag{1}$$

其中，Sim(q, l) 表示查询 q 与候选逻辑链 l之间的相似度分数，该分数通过双编码器架构计算。softmax 分母促使模型为黄金逻辑链赋予最高分数。

每个查询 q 都与唯一的逻辑链 l相关联，使得检索到的逻辑链 l^+ 能够蕴含支持性文档以及回答 q所需的推理步骤。这种对齐使检索器能够支持两种下游任务。问答到逻辑链检索：给定一个临床问题 q，检索与之对应、体现基于指南进行推理的逻辑链 l^+ 。指南到逻辑链检索：给定一段文档摘录 d，检索适用于诊断或治疗推断的推理路径 l^+ 。通过对这两种任务进行联合训练，检索器能够学习问题、推理链和医学文档之间具有泛化能力的对齐，从而支持在多种输入场景下进行结构化检索。

A.4.2. 评估结果

我们比较了所提出的逻辑链检索器与强基线 ColBERT 检索器在 top-k 证据选择任务上的检索准确率。对于测试数据集，我们基于 PubMed 语料库构建问答对进行测试。如表 6所示，在所有 k取值下，逻辑链检索器的表现始终优于 ColBERT。值得注意的是，其 top-5 准确率达到 73.27%，而 ColBERT 为 62.33%。随着 k 增大，这一性能差距进一步扩大，达到 81.79%

表 *6.* 逻辑链检索器与 ColBERT 检索器的 Top-k 检索准确率（%）对比。

逻辑链检索器	73.27	77.49	80.13	81.79
ColBERT 检索器	62.33	65.12	67.48	70.51

在 top-20 检索时，这表明在检索过程中融入结构化推理信号是有效的。这些结果表明，显式建模逻辑链不仅能提升与查询意图的对齐程度，还有助于检索到更多相关且具有解释性的证据。

A.5. 问答生成示例

给定一个逻辑链，我们会自动生成问答对：将最后一步设为答案，并把前面的推理路径改写成问题。这样的对齐确保问题在逻辑上由链中的步骤推出，而答案则是推理过程自然得出的结论。

例如，给定以下输入：

文档：“顺铂与DNA结合并诱导交联，从而导致细胞凋亡。”**逻辑链**：“顺铂 $\rightarrow$ DNA交联 $\rightarrow$阻断复制 $\rightarrow$ 细胞凋亡”

我们生成：

问题：顺铂的治疗机制是什么？**答案**：DNA交联

这个问答对既反映了事实内容，也体现了推理结构，因此可用于对齐任务或作为监督微调的目标。必要时，我们会采用受控变体（例如被动语态、问题句式）来增加语言多样性，同时保留底层逻辑。

A.6. 评估提示设计

评估封闭式问答时，我们只需计算准确率。然而，许多开放式问答任务（例如 Amboss数据集中的诊断推理问题）还带来了额外的挑战。尽管存在多种衡量文本相似度的方法，例如 F1 或 ROUGE，但这些方法在医学领域都有明显的局限性。因此，我们提出了一套非常严格的评估流程，采用两项评估指标：USMLE事实性分数和GPT-4o分数。对于 GPT-4o分数，直接让 GPT-4o 为答案评分往往效果不佳，因为 GPT-4o 倾向于偏好符合其语言风格的答案，而这可能与我们预期的判据不一致。因此，我们引入了一个评分框架，由医生设计，从三个方面评估模型的细粒度诊断能力：**关键点**、**推理**和**证据**：

- **关键点** 评估模型的答案是否包含标准答案中的关键要素。
- **推理** 评估模型答案中的诊断推理是否正确、是否遵循与标准答案相同的步骤，以及是否遗漏了任何关键步骤。
- **证据** 用于评估模型的答案是否提供了支持其结论或诊断推理的关键证据。

最后，将计算平均分数，以代表答案的整体质量。为进一步降低语言风格对 GPT-4 评分的影响，我们提出使用 GPT-4 修订所有模型生成的答案，确保所有输出都符合 GPT-4 自身的风格分布。在修订过程中，GPT-4 只能看到模型的答案，无法访问任何其他信息。

评分时，GPT-4 将根据标准答案生成自己的**关键点**、**推理**和**证据** 摘要。在为这些方面打分时，GPT-4 将不再查看原始答案，而只参考其总结的**关键点**、**推理**和**证据**。

表 7. 评估流程提示示例模板。

系统提示	以下文本包含医学信息。你的任务是创建 基于所呈现医学知识的规则，确保每条规则 全面体现证据文本中的信息，同时不 不遗漏任何重要细节。
清晰区分	每条规则都应清楚说明哪些项目属于条件、患者 特征或前提条件（例如特定诊断、临床 发现或情形），哪些属于结果、措施或 建议。始终将条件列在规则开头，随后 由推荐的措施或结果。
多条规则用于 复杂证据	如果一条证据陈述包含多个不同方面，请创建单独的 规则，针对每个部分。这样可确保每个概念得到体现 得充分且清晰，防止不相关的想法被合并到一个 单条规则。根据需要，为每条不同的规则重复使用证据。
并列项 和逻辑 结构	在每条规则中按逻辑顺序排列各项。- 对于具有自然 顺序的项目（基于现实世界的逻辑，例如患者特征 -> diag- 断 -> 治疗建议），请遵循此顺序。- 如果各项 是并列关系，没有严格的先后顺序，则应在 规则中将它们置于同一层级，不强加顺序，例如“PET获益 -> 诊断， 分期、治疗反应”，前提是它们都同等适用。
现实逻辑 顺序	超越句子结构**：在适用时，遵循自然的顺序 根据现实世界中的关系，而非句子顺序。首选 顺序通常如下：- **条件或要求** （例如，患者特征或具体标准）- **诊断检查**（例如， 已进行的检查）- **发现或结果**（例如，诊断检查的 结果）- **治疗建议或措施**（例如，拟议的 根据所见采取的治疗）- **随访或监测计划**（如 适用）请按此逻辑顺序呈现这些要素，无论它们在原文中的排列顺序如何。 它们在原文中的排列方式。
重建 指南	确保生成的证据和规则集合涵盖该指南 全面呈现，以便准确重构指南 中的这些部分。每条证据都应尽可能贴近 原始指南文本，并在可行时使用精确措辞。
包含详细信息	呈现证据中的每个重要术语、短语或概念。 每个规则项都应尽可能简洁，同时完整呈现 所有关键细节。避免不必要的冗余，但确保包含所有必要 信息。使用清晰、直接的语言，避免过度 使用“如果”或“当.....时”等条件句。每条规则在逻辑上都应 完整且自成一体，传达信息的完整上下文。 tion.

表 *8.* 评估流程提示示例模板（续）。

输出格式	**需包含的关键词**：关键词 **医学文本**：医学文本 **医学路径**：医学路径 **示例规则格式**：- **正确**：- **证据**："患者出现腹痛，随后进行了 CT扫描，结果显示有脓肿。治疗包括抗生素。" - **规则**：腹痛 -> CT扫描 -> 检测到脓肿 -> 治疗：抗生素。- **错误**（完全依据句子顺序）：- **证据**："患者出现腹痛，随后进行了 CT扫描，结果显示有脓肿。治疗包括抗生素。" - **规则**：腹痛 -> 检测到脓肿 -> CT扫描 -> 治疗：抗生素。 **返回格式**：1. 证据：规则：2. 证据：规则：仅按指定输出格式返回，不要添加其他评论或解释。

B. 案例研究

表*11.* 人工标注样例。

问题： 一名67岁男性，患有膀胱移行细胞癌，因 耳鸣2天，前来就诊。一周前，他接受了首个疗程的 新辅助化疗。纯音测听显示感音神经性听力损失， 听力损失程度为45 dB。导致该患者出现这些症状的药物，预期最有益的作用是 可能是由以下哪项作用导致的？A：抑制蛋白酶体，B：微管过度稳定， C：自由基生成，D：DNA交联
Phi-4: 要确定导致患者出现这些症状的药物预期具有何种有益作用， 我们需要确定最可能导致感音神经性听力损失的药物，并了解 其作用机制。1. 患者患有膀胱移行细胞癌，并接受了 新辅助化疗。2. 膀胱癌常用的一种化疗药物是顺铂。 3. 已知顺铂会引起耳毒性，可能表现为感音神经性听力损失。4 顺铂会导致自由基生成并引起听力损失，因此最终答案应为 C。
Qwen-2.5-14B-Instruct 膀胱移行细胞癌的化疗通常包括顺铂等铂类药物 。顺铂以其耳毒性而闻名，可能 导致感音神经性听力损失。-> 顺铂的作用机制涉及自由基生成 自由基可能会损伤细胞成分，包括听觉系统中的细胞成分。-> 因此，自由基生成最有可能是导致患者症状的作用。 ->正确答案是 C
TARSE： **小批量微调：** **预训练数据：** MDMA：3,4-亚甲基二氧甲基苯丙胺（MDMA），俗称摇头丸（E）或莫莉，是一种主要用于娱乐用途的精神活性药物。其预期效果包括感官变化、精力增强、共情以及愉悦感。口服后，药效会在 30 至 45 分钟内开始，并持续 3 至 6 小时..... **微调数据：** 场景：经验丰富的泌尿科医师史密斯医生正准备为一名45岁男性患者实施手术，以修复前尿道狭窄。患者得知，一名外科住院医师将参与手术，作为其培训的一部分。患者担心住院医师的参与可能会对手术产生.....

推理： 1. 膀胱移行细胞癌 ->新辅助化疗 -> 导致感音神经性听力损失的耳毒性药物 -> 该药物的正确作用是使微管过度稳定 2. 膀胱移行细胞癌 -> 新辅助化疗 -> 耳毒性副作用 -> 感音神经性听力损失 -> 由自由基生成所致 3. 膀胱移行细胞癌 -> 新辅助化疗 -> 机制为DNA交联 -> 导致耳毒性的化疗药物 -> 出现感音神经性听力损失 -> 耳毒性机制涉及自由基生成 -> 答案：E

检索以验证： 1. 顺铂是一种化疗药物 → 会导致耳毒性（听力损失） →原因是耳蜗中活性氧（活性氧）蓄积 → 活性氧引起氧化应激并损伤耳蜗毛细胞 → 抗氧化剂中和活性氧并减少耳蜗损伤→ 因此，抗氧化剂有助于预防顺铂引起的听力损失 2. 化疗药物-> DNA交联->治疗

最终答案： D

医生评论： 该临床推理任务要求区分药物的治疗机制与其副作用特征。虽然Phi-4和Qwen-2.5-14B等模型正确识别出顺铂是致病药物，但它们将药物的作用机制错误地归因于其耳毒性副作用（自由基生成）。相比之下，TARSE正确区分了治疗机制（DNA交联）与副作用，展现出一条结构化的推理路径，与人工临床逻辑相符。

表 *9.* 评估流程提示示例模板。

提取关键点	根据问题和答案，总结你认为标准答案中最关键的十个要点。请将回答以以下格式返回：{1.2.3....} 以下是问题：{question} 以下是答案：{answer} 请勿提供任何其他信息-息 .
关键点	1. 多焦视网膜电图（ERG）显示右眼整个黄斑区域的信号减弱，确诊为急性区域性隐匿性外层视网膜病变。2. 急性区域性隐匿性外层视网膜病变（AZOOR）最早由 Gass 在 1993 年.....
提取诊断推理	根据问题和答案，请详细总结标准答案中的诊断推理。请按以下格式返回回复：{1.2.3....} 问题如下：{question} 这里答案是：{answer} 请勿提供任何其他信息。
诊断推理	1. 患者是一名7岁男孩，自出生以来，左下颈部有一个缓慢生长、无症状的肿块。2. 体格检查显示一个黄色、驼峰状肿块，表面有毛发，质地类似软骨左侧胸锁乳突肌附近的一致性.....
提取证据	请根据问题和答案，提供一份详细的证据列表，该列表由正确答案提出。请按以下格式返回回复：{1.2.3....} 问题如下：{question} 答案如下：答案：{answer} 请勿提供任何其他信息。
证据	1. 自出生起，左下颈部便有一个缓慢生长、无症状的肿块。2. 体格检查发现一个黄色、驼峰状肿块，表面有毛发，质地如软骨。3. 超声检查显示一个低回声、无血管、隆起的结节，伴有无回声管状结构。4. 低回声、无血管、隆起的结节，伴有无回声管状结构。4. 磁共振成像显示一个隆起性结节，伴有弥漫性.....
要点得分	请担任美国执业医师资格考试评估员，您的职责是评估并比较医学生的解释与提供的目标答案。请先仔细审阅提供的目标答案，再开始评估。然后，根据以下具体标准，确定学生答案的分数。请判断医学生的答案是否包含这些关键点（或其他相关要点。但要点的数量必须完整）。例如，标准答案有 10 个关键点，如果学生答案包含一个关键点，他将得到 0.5 分（如果答案包含 5 个要点，则应为 2.5）。医学生的答案：{answer} 关键点：{Key Point} 请仅返回一个浮点数（从 0 到 5）。你应该逐一检查每个要点（不应根据语言风格进行判断例如流利度等。仅根据学生的答案是否包含正确、相关且完整的关键点来判断）。 不要生成任何其他信息。

表*10.* 医疗指南、信息和语料统计来源

来源/语料库全称		指南/#文档	词语/#片段	受众/领域	国家	已发布
CCO	安大略癌症护理中心	87	199K	医生	加拿大	Yes
CDC	疾病控制与预防中心	621	6.7M	医生	USA	Yes
CMA	加拿大医学会	431	1.7M	医生	加拿大	Yes
ICRC	红十字国际委员会	49	1.2M	医生	国际	Yes
NICE	英国国家卫生与保健卓越研究所	1656	8.1M	医生	UK	Yes
PubMed	PubMed	1627	1080万	医生	Mix	Yes
SPOR	患者导向研究战略	217	1.1M	医生	加拿大	Yes
WHO	世界卫生组织	223	3.1M	Both	国际	Yes
WikiDoc	WikiDoc	33058	34M	Both	国际	Yes
StatPearls	StatPearls	9.3k	301.2k	Both	国际	Yes
教科书	教科书	18	125.8k	Both	国际	Yes
维基百科	维基百科	6.5M	29.9M	Both	国际	Yes
MedCorp	MedCorp	30.4M	54.2M	Both	国际	Yes