评估用于罕见病诊断的多智能体大语言模型架构

Ahmed Almasoud

人工智能与数据分析（*AIDA*）实验室，计算与信息科学学院，*Prince Sultan University*，沙特阿拉伯利雅得。

电子邮箱： *almasoud@psu.edu.sa*

摘要

尽管大语言模型是有效的诊断工具，但多智能体拓扑结构对诊断准确率的影响仍缺乏充分研究。本研究评估了四种智能体拓扑结构：对照组（单智能体）、层级式、对抗式和协作式，涵盖 33 种罕见病类别中的 302 个病例。我们提出一种推理差距指标，用于量化内部知识检索与最终诊断准确率之间的差异。结果显示，层级式拓扑结构（准确率为 50.0%）略优于协作式（49.8%）和对照组（48.5%）配置。相比之下，对抗式模型显著降低了性能（27.3%），并呈现出巨大的推理差距：有效诊断因人为怀疑而遭到否定。在所有架构中，模型在过敏性疾病和毒性效应类别中的表现最佳，在心脏畸形和呼吸系统病例中的表现则最差。值得注意的是，尽管单智能体基线总体上表现稳健，但所有多智能体系统（包括对抗式模型）在骨骼和胸部疾病类别中的准确率均更高。这些发现表明，提升系统复杂性并不能保证推理能力更强，支持转向动态拓扑选择。

关键词： 多智能体大语言模型、罕见病、GPT-5.1、疾病诊断

1. 引言

将大语言模型（LLM）整合到临床工作流程中，标志着诊断辅助领域的一项重大转变。近期荟萃分析表明，大语言模型辅助能够提升不同医学领域和职业阶段医生的诊断准确率 [1]。然而，将这些模型应用于罕见且复杂的病理仍面临诸多挑战。基准研究显示，通用模型虽然在常见临床表现方面表现良好，但在应对罕见遗传病时，其精确率往往不及传统生物信息学工具 [2], ，一些最先进模型在叙述性罕见病任务中的准确率甚至低至 16.5% [3]。

为缓解单模型推理固有的幻觉和推理失误，研究人员正越来越多地转向多智能体系统（MAS）。这类架构试图复现临床实践共同体，例如多学科团队（MDT）或分层监督结构 [4]。现有研究，如 MedChat [5] 和 AI Hospital [6]，表明将任务分配给特定角色智能体能够提高可靠性和模拟保真度 [7]。此外，通用领域的理论进展表明，多智能体辩论可以通过社会纠错提升事实准确性并减少错误 [8]。

尽管取得了这些进展，特定智能体拓扑结构对诊断精度的影响仍未得到充分探究。本研究通过评估四种不同的智能体架构——对照、层级式、对抗式和协作式——来弥补这一空白，评估对象为 Chen等人整理的302例罕见病病例 [9]。我们还引入了一项新指标——推理差距，用于量化内部知识检索与最终裁定之间的差异。

2. 相关工作

关于人工智能驱动的医学诊断的研究，已从评估单个大语言模型发展到设计复杂的智能体工作流。Dinc等人 [10] 观察到，尽管 Claude 3.7 等先进模型在常见病例上的准确率近乎完美，但在复杂场景中的表现差异显著。这一局限在罕见病领域尤为突出；Reese等人 [2] 证明，通用型大语言模型的诊断排序性能无法媲美 Exomiser 等表型专用工具。同样，Gupta等人 [3] 使用源自《豪斯医生》的数据集表明，即使是前沿模型，在应对长尾医学异常时也困难重重，因此需要采用超越简单零样本提示的架构。

2.1 层级式与基于角色的架构

为解决单智能体的不足，近期研究强调智能体监督和层级委派。Kim 等人 [11] 提出了分层智能体监督（TAO），模拟医院层级结构（例如护士—医生—专科医生）。研究结果表明，自适应分层监督可以吸收多达24%的个体错误，这说明结构层级能够充当噪声过滤器。这与Zuo等人 [12], 的研究一致，其 KG4Diagnosis 框架采用全科医生智能体进行分诊，并由专科代理开展深入分析，同时借助知识图谱加以增强。

角色扮演智能体的效用进一步得到了Wang, Z.等人[13], 的支持，他们提出了MedAgent-Pro。通过将诊断解耦为疾病层面规划和患者层面推理，该研究表明，结构化的逐步工作流程可减少经验性幻觉，这一发现得到了Peng等人[14], 的强化，他们实施了推理树框架。特定临床领域也从该方法中受益；例如，Ji等人[15] 为中轴型脊柱关节炎开发了SpAgents，其中不同的规划智能体、数据智能体和医生智能体相互协作

从而表现优于初级风湿病学家。同样，研究人员利用混合智能体来模拟肿瘤学 [16] 和癌症护理 [17], 中的肿瘤讨论组，证实多模态、多视角系统比单体模型更符合临床共识。

2.2 辩论、协作与聚合机制

多智能体设计中的一个关键差异在于共识形成方式。Chen等人 [9]证明，受多学科团队讨论启发的多智能体对话（MAC）框架，在罕见病诊断方面优于思维链（CoT）和自一致性方法。Thakrar等人 [18], 的研究进一步支持了这种协作方法：他们发现，结合类似同行会诊视角的临床启发式推理层，比单独进行微调能取得更高的准确率。

与此同时，Du等人 [8] 认为，要求智能体批评并辩论回应，有助于提升通用领域中的事实准确性和策略推理。然而，将这种辩论机制应用于医学领域仍存在争议。Feng, Y.等人 [19] 利用强化学习优化了多轮对话中的提问策略，但单纯辩论能否有效消解医学歧义，仍是一个尚待解答的问题。此外，如何聚合这些多智能体输出也面临自身的挑战。Ai等人 [20] 提出了“逆惊人流行度”等算法，以缓解简单多数投票的局限性；这表明，必须通过数学方法考量智能体之间的相关性与异质性，才能确保集体决策可靠。

我们的研究以这些基础性研究为基础，直接比较临床文献中常见的协作式和层级式模型 [9, 11] 与通用人工智能研究中提出的对抗式辩论模型[8], ，并将其置于罕见病诊断的具体情境中进行考察。

3. 方法

图1a提供了研究方法的示意图概览。所有诊断实验均使用通过OpenAI API访问的GPT-5.1大语言模型进行。

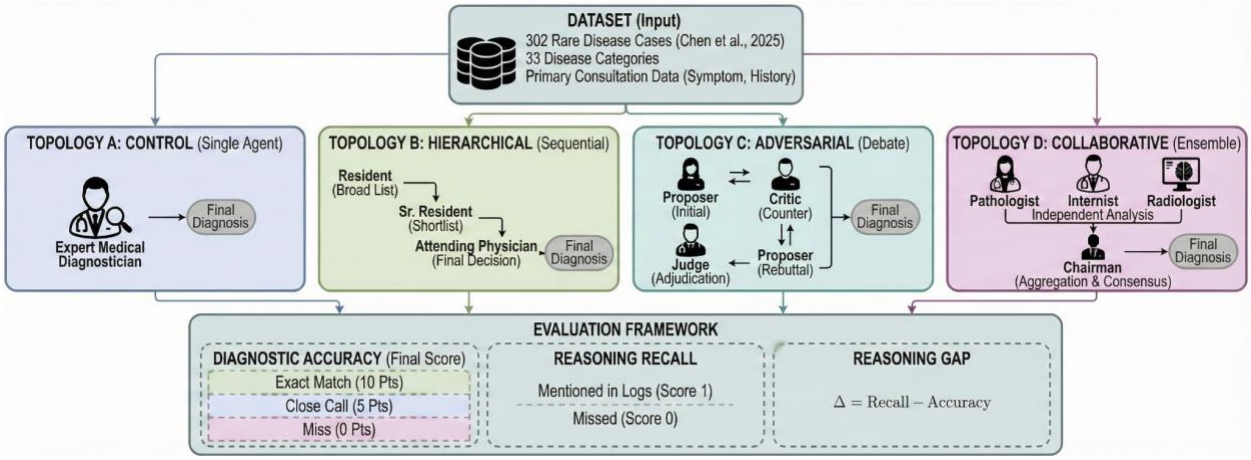

图1a. 实验方法示意图。

3.1 数据集

本研究使用由Chen等人整理的、公开可用的302例罕见病病例数据集[9]。这些病例来源于真实世界的临床报告，涵盖来自Orphanet数据库的33个不同疾病类别。每条记录代表一个初诊场景，提供患者人口统计信息、体格检查、病史和初步检查结果，以复现首次临床接诊时的不确定性。该数据集包含原始病例报告中已确认的最终诊断，作为我们评估的真实标签。

3.2 多智能体架构

为评估不同推理结构对诊断准确率的影响，我们设计了四种不同的拓扑配置。这些配置从标准的单智能体基线，到模拟不同临床工作流程的复杂多智能体系统，涵盖了不同复杂程度。

3.2.1 对照组（单智能体）

对照拓扑结构代表了大语言模型的基线表现。它采用直接的零样本提示策略，由一个扮演“医学诊断专家”的单智能体分析患者数据，并输出最可能的单一诊断，不经过中间推理步骤，也不接收外部反馈。该设置作为基准，用于与多智能体架构进行比较。

3.2.2 层级式（顺序式）

这一拓扑结构模拟了分层式医院工作流程，旨在将广泛的鉴别诊断筛选为具体结论。如**方案1.1**（图1b）所详述，我们将该系统构建为三级漏斗。流程始于一名**住院医师**智能体，其分析病例并生成包含三项潜在诊断的列表。随后，该输出被传递给**高年资住院医师**，由其审阅住院医师的列表，依据症状分析排除一个不太可能的候选诊断，并

筛选出前两项诊断。随后，一名 **主治医师** 从高年资住院医师的候选名单中选出最可能的单一诊断，从而做出最终决定。

1.1. Hierarchical Diagnosis Protocol

Input: Patient Case (D), Language Model (M)
Output: Final Diagnosis (Y)

1: **Resident:** $L_{res} \leftarrow M$("List 3 potential diagnoses for D")
2: **Senior Resident:** $L_{short} \leftarrow M$("Filter L_{res} to top 2 candidates given D")
3: **Attending Physician:** $Y \leftarrow M$("Select single most likely diagnosis from L_{short}")
4: **return** Y

图1b. 分层诊断方案

3.2.3 对抗式（辩论）

这种拓扑结构引入冲突驱动的推理，以减轻确认偏误。如**方案 1.2**（图1c）所述，该系统强制**提议者**与**批评者**之间进行结构化辩论。与顺序模型不同，这种拓扑结构明确指示批评者代理寻找矛盾证据，而不论初始诊断的强度如何。这确保最终裁决者（“裁判”）面对的是经过压力测试的论证，而非单一且无人质疑的观点。裁决者随后审阅整个交流过程（提案、批评意见、反驳），以裁定最终诊断。

1.2. Adversarial Debate Protocol

Input: Patient Case (D), Language Model (M)
Output: Final Diagnosis (Y)

1: **Proposer Agent:** $H_0 \leftarrow M$("Propose diagnosis for D")
2: **Critic Agent:** $H_1 \leftarrow M$("Find contradiction in H_0 given D")
3: **Proposer Agent** – Rebuttal)**:** $H_{final} \leftarrow M$("Defend H_0 against H_1")
4: **Judge:** $Y \leftarrow M$("Adjudicate H_0, H_1, H_{final}")
5: **return** Y

图1c. 对抗式辩论方案

3.2.4 协作式（集成）

这种拓扑结构通过模拟多学科团队（MDT）的工作方式，汇集众人智慧。如**方案 1.3**（图 1d）所示，三位专科代理——病理学家、内科医生和放射科医生——同时且独立地分析病例。随后，主席代理结合病例数据与这三位专家各自独立的意见，确定最终的共识诊断。

1.3. Collaborative Ensemble Protocol

Input: Patient Case (D), Language Model (M)
Output: Final Diagnosis (Y)

Parallel Analysis:
  $O_{path} \leftarrow M(\text{“Pathologist: Analyze } D\text{”})$
  $O_{int} \leftarrow M(\text{“Internist: Analyze } D\text{”})$
  $O_{rad} \leftarrow M(\text{“Radiologist: Analyze } D\text{”})$
Aggregation:
  Chairman: $Y \leftarrow M(\text{“Synthesize } \{O_{path}, O_{int}, O_{rad}\} \text{ into consensus”})$
return Y

图 1d. 协作集成协议

3.3 评估框架

为严格评估每种拓扑结构的表现，我们设计并实现了一套评估系统，用于衡量最终输出的准确率以及底层推理过程的广度。

3.3.1 诊断准确度评分

我们采用以大语言模型为评审的方法评估模型，由 GPT-5.1 为每项诊断评分。自动裁定器（大语言模型）使用三级评分标准，将模型的预测与真实标签进行比较，区分完全匹配（10 分）、临床相关鉴别诊断（5 分）和完全未命中（0 分）。完整的评分标准逻辑见

图 2。

为便于比较分析，最终的**诊断准确度评分**按数据集上的平均分计算，并归一化为百分比：

$$\text{Diagnostic Accuracy Score} = \left(\frac{1}{N} \sum_{i=1}^{N} S_i \right) \times 10$$

其中，N=302 表示病例总数，Si∈{0,5,10}表示分配给第 i 个病例的分数。通过对相关鉴别诊断给予部分分，该指标反映了严格二元正确性之外的诊断效用。

尽管近期研究表明，大语言模型在开放式评估中可能存在自我偏好偏差[21], ，但我们通过使用真实标签和评分标准直接缓解了这一问题（图 2）。这使评估者依据客观诊断标准进行评判，而非主观偏好。

此外，由于同一裁决者（GPT-5.1）评估了所有四种拓扑结构（对照、层级式、对抗式、协作式），任何残余偏差在各实验条件下都保持一致，从而确保了比较分析的有效性。

Evaluation Rubric (LLM-as-a-Judge)

System Instruction: You are a Medical Adjudicator. Compare the Prediction vs Ground Truth.

SCORING RULES:

[10 POINTS] Exact Match
The diagnosis is identical to the ground truth or is a recognized direct synonym (e.g., "Stein-Leventhal Syndrome" == "PCOS"). Ignore minor formatting differences.

[5 POINTS] Close Call / Differential
The predicted disease is in the exact same family or is a direct subtype/parent of the truth.
Example: Truth="Duchenne Muscular Dystrophy", Prediction="Muscular Dystrophy" (Generic).
Example: Truth="Viral Meningitis", Prediction="Meningitis" (Unspecified).

[0 POINTS] Miss
The prediction involves a completely unrelated pathology, organ system, or etiology.

图 2. 评估标准

3.3.2 推理差距

为量化决策过程的效率，我们引入了一项名为**推理差距**的新指标。该指标区分了检索失败（模型始终未能识别出正确的病理）和裁决失败（模型在审议过程中否定了正确诊断）。

首先，我们通过扫描完整的交互日志（例如住院医师的初始列表或提议者的论据），计算基线**推理召回率**得分，以确定是否曾明确考虑过正确诊断，而不论最终输出如何。具体而言，若明确出现真实诊断（通过**严格关键词匹配**），则赋予 10 分；若未出现，则赋予 0 分。随后，对这些二元得分取平均值，并按每种拓扑结构。归一化为百分比（0–100%）。

我们将推理差距定义为这一潜在表现（召回率）与实际表现（准确率）之间的差值：

$$\Delta = Reasoning\ Recall\ (\%) - Diagnostic\ Accuracy\ (\%)$$

这一派生指标即为拒绝错误。较大的正差距表明，拓扑结构的知识库中包含正确诊断，但未能将其选为最终答案，例如受到说服性干扰（在对抗式模型中）或过度过滤（在层级模型中）的影响。反之，差距极小则表明决策边界十分高效，能够将识别出的模式正确转化为最终诊断。

3.3.3 亚组分析

为评估这些拓扑结构在不同医学领域中的稳健性，我们依据**疾病类型**进行了分层分析。我们汇总了 33 个不同临床类别（例如肿瘤性、感染性、遗传学、免疫学）的表现得分，以找出每种多智能体架构固有的特定领域优势或劣势。

4. 结果与讨论

4.1 总体诊断性能

四种拓扑架构的比较结果表明，系统复杂性与诊断成功率之间呈非线性关系。如**表1**所总结并在**图3**中直观呈现的那样，智能体的结构安排既影响正确医学知识的检索（推理召回率），也影响将这些知识最终用于得出正确诊断的能力（准确率）。

表1. 各拓扑结构的性能指标汇总

拓扑结构	**诊断 准确率（%）**	**推理召回率（%）**	**推理差距（Δ）**
对照组（基线）	48.5%	N/A	N/A
层级式	**50.0%**	**54.0%**	4.0
协作式	49.8%	51.3%	1.5
对抗式	27.3%	44.0%	**16.7**

层级式拓扑结构表现最佳，诊断准确率达到 **50.0%**。与单智能体**对照**基线（48.5%）相比，这一结果虽提升幅度不大，但表现稳定（提高 +1.5%）。值得注意的是，层级模型的**推理召回率（54.0%）**也最高，这表明，先由住院医师智能体给出广泛的鉴别诊断，再通过高级审核加以细化的顺序漏斗结构，是找出正确真实标签的最有效策略。这与分层智能体监督的研究结果一致：分层监督能够有效吸收单个智能体的错误（**Kim 等人，2025**）。

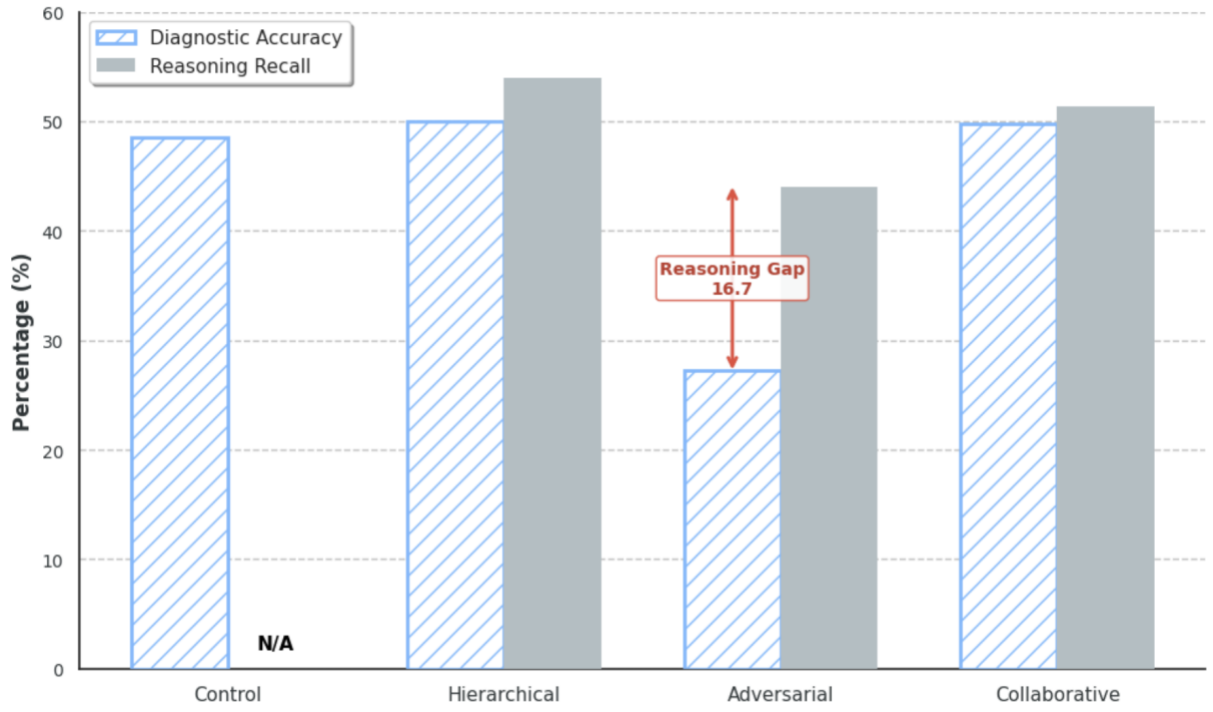

图3. 不同拓扑结构下的性能

注：由于对照拓扑结构仅直接输出最终诊断，不会生成多智能体拓扑结构所产生的中间交互日志，因此其推理召回率标记为不适用。

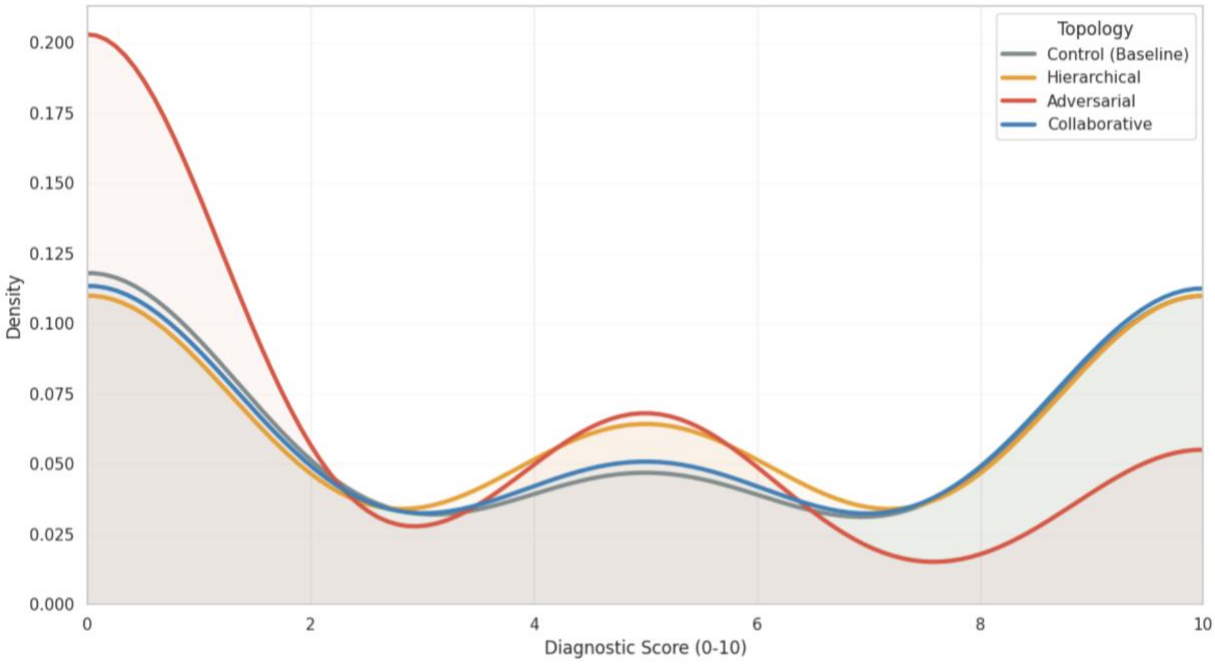

图 4. 多智能体各拓扑结构及对照组（单智能体）的诊断评分概率密度

协作拓扑（集成）在最终输出方面的表现与层级模型相当，诊断准确率为 **49.8%**，召回率为 **51.3%**。然而，最引人注目的发现是，**对抗式拓扑**的表现大幅下降。尽管其推理召回率达到 **44.0%**，最终诊断准确率却骤降至 **27.3%**。如概率密度图（**图 4**）所示，对抗式模型的评分分布明显向低分端偏移。这表明其决策边界存在系统性失效：在辩论过程中，提议者或批评者代理往往能够识别出正确的疾病，但裁决者代理却常常受到反驳论点的影响，最终舍弃有效诊断，转而选择看似合理却错误的替代诊断。

这些结果对“增加智能体交互必然能提升推理能力”这一假设提出了质疑。层级模型和协作模型增加智能体后，系统的表现略有提升，优于单智能体；但对抗式模型引入的冲突造成了破坏性干扰，严重削弱了系统的精确率。

4.2 推理差距分析

我们分析了**推理差距**（Δ），将其定义为**推理召回率**（知识检索）与**诊断准确率**（最终决策正确性）之间的定量差异。层级模型的差距保持在适中的 4.0，表明其多阶段审查流程（住院医师 → 高级医师 → 主治医师）发挥了有效的临床筛选机制作用。同样，协作拓扑的差距较低，为 1.5，说明主席代理很少推翻专家共识。

最显著的异常出现在对抗式拓扑中，其推理差距高达16.7。这一差距表明，裁决阶段出现了系统性失效。对交互日志的定性分析表明，强制批评这一约束经常掩盖正确的诊断。由于批评者代理被要求无论初始诊断的可信度如何，都必须寻找矛盾证据，因此它经常生成看似合理但最终无关的干扰信息。在提议者正确识别出疾病的情况下，批评者被迫采取的怀疑态度引入了人为怀疑；相反，当批评者正确识别出漏诊时，提议者的防御性反驳往往会说服裁判坚持错误的初始假设。这表明，尽管辩论能够促进创意生成，但无条件怀疑可能会降低罕见病诊断所需的精确率。

4.3 特定领域表现（亚组分析）

我们分析了33种疾病类别中的性能表现，如表2所示。这项分析通过性能增益与对照（图6）以及诊断难度（图5）进行可视化，突出了由疾病复杂性驱动的不同行为模式。

表2. 根据评估标准，按疾病类别划分的各拓扑诊断评分

疾病类型	对照	层级式	对抗式	协作式
腹部外科	2.50	2.50	1.25	2.50
过敏性	10.00	9.00	4.00	10.00
Bone	3.85	5.77	4.62	5.38
心脏	6.25	6.25	3.75	6.25
心脏畸形	1.25	1.25	1.25	1.25
循环系统	6.25	5.63	4.38	5.00
胚胎发生	2.14	3.57	1.43	3.57
内分泌	5.50	7.00	4.00	6.00
胃肠病学	6.25	6.25	1.88	8.13
遗传学	3.67	4.67	2.33	3.33
妇科与产科	5.56	4.44	2.22	5.56
血液学	5.00	5.00	5.00	4.00
肝脏	8.75	7.50	3.13	8.75
免疫学	5.00	5.50	2.50	4.00
先天性代谢	3.64	3.64	2.73	3.64
感染性	3.75	5.63	1.25	5.00
不孕症	5.63	5.63	1.88	5.63
肿瘤性	4.17	2.50	1.67	3.33
神经系统	4.23	4.23	2.69	4.62
牙科学的	5.00	7.14	3.57	6.43

眼科	5.00	4.09	2.27	4.55
耳鼻喉科	6.00	5.00	3.00	6.00
肾脏	3.33	3.33	2.78	3.89
呼吸系统	1.43	0.71	1.43	5.00
风湿性疾病	5.00	6.00	3.00	5.50
儿童风湿	10.00	8.33	5.00	8.33
Skin	2.73	5.00	1.82	3.18
颌面外科	3.33	5.00	2.50	4.17
畸形学	6.67	5.83	2.50	5.00
胸外科	2.50	5.00	4.38	3.13
毒性效应	8.33	8.33	3.33	8.33
移植相关	5.91	4.09	3.18	4.55
泌尿生殖	5.63	7.50	3.13	8.13

病例的内在难度差异显著。如图5所示，过敏性疾病（平均评分为 9.7，不包括对抗式拓扑结构）、毒性效应（8.3）和肝脏疾病（8.3）等类别对大语言模型而言较易解决。相反，心脏畸形（1.2）和腹部外科病例（2.5）等结构歧义较大的类别，无论采用何种架构拓扑结构，都难以诊断。不过，呼吸系统疾病的表现出现了显著差异。对照模型和层级模型处理这类病例时均困难重重（评分分别为 1.43 和 0.71），而协作拓扑结构取得了 5.00 的评分。这一提升表明，多视角集成能够独特地综合呼吸病理学中复杂且相互交叠的症状。

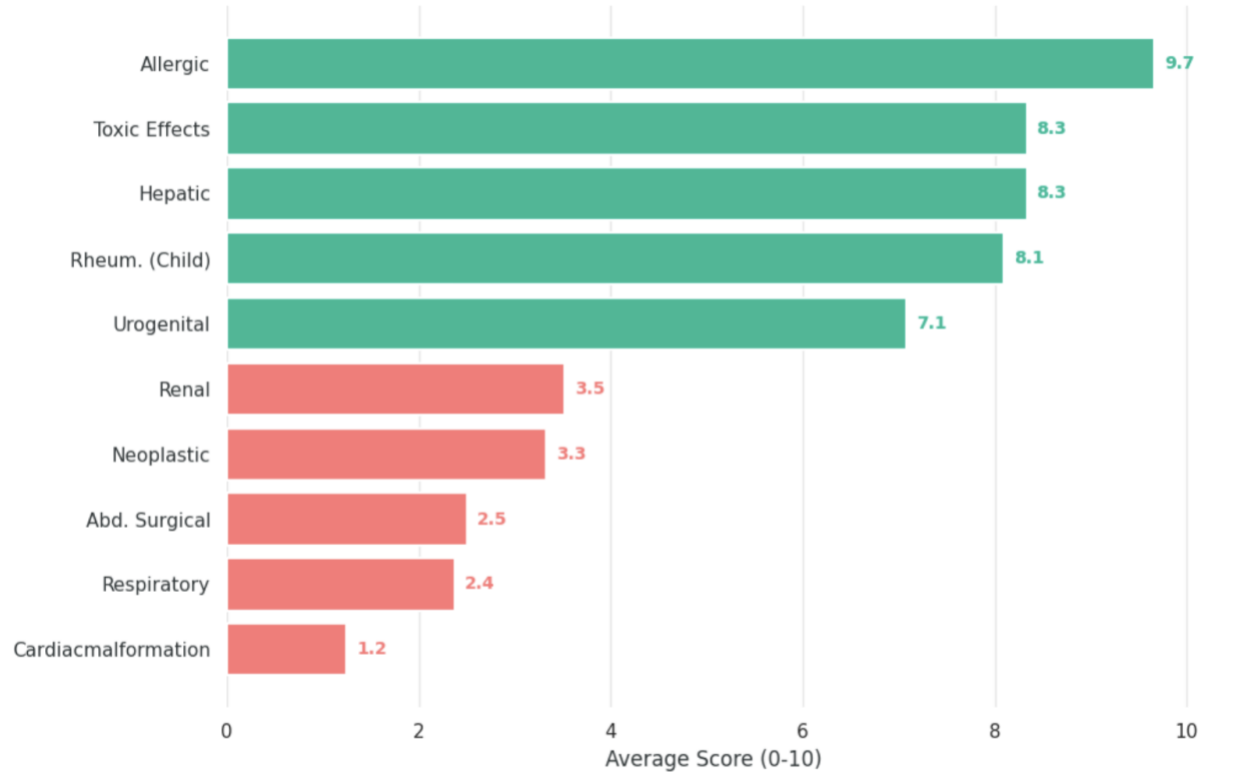

图5. 各疾病类别的平均诊断性能（不包括对抗式拓扑结构）

注：评分表示对照、层级式和协作拓扑结构的平均诊断准确度评分（*0–10*）。此图未纳入对抗式拓扑结构，以突出疾病的内在难度，避免基于辩论的模型中观察到的破坏性干扰影响结果。

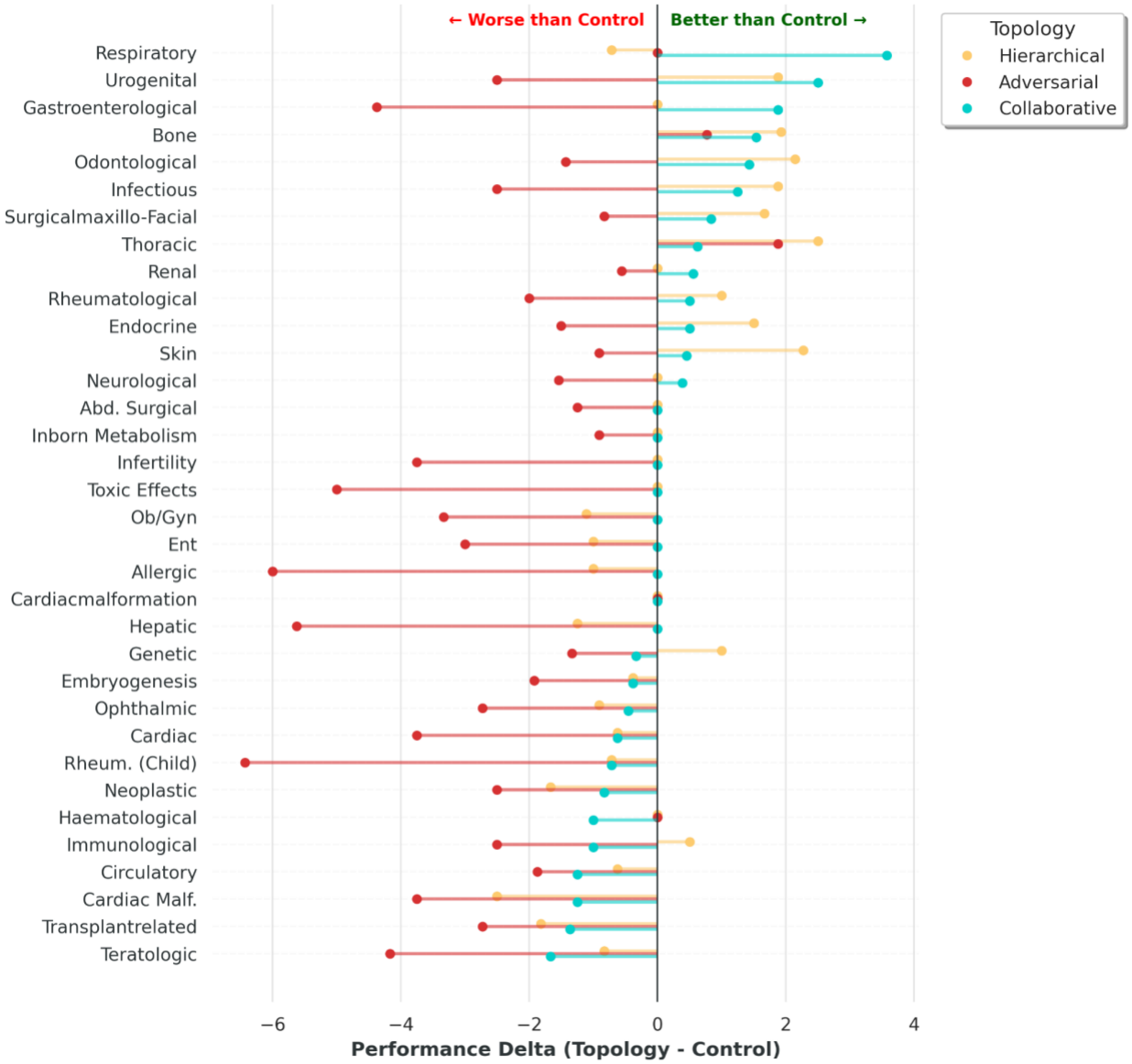

图6. 各拓扑结构与对照（基线）的性能比较

图6展示了各多智能体架构相对于对照基线的表现（基线由 x=0 处的中心轴表示）。一个引人注目的发现是，单智能体对照组表现稳健。在一些复杂类别中，尤其是畸形学、移植相关和循环系统，对照智能体的表现优于多智能体集成，这些集成位于基线左侧。这对“增加智能体总能改善结果”这一假设提出了挑战。协作拓扑在某些病例中确实取得了显著提升，例如呼吸系统（+3.6）和泌尿生殖系统疾病病例（涉及重叠的器官系统），因为综合不同症状是关键。不过，它在某些类别中未能胜过单智能体，这也引发了对其效率的合理质疑。

与之形成鲜明对比的是，对抗式拓扑始终位于基线左侧，表明其表现出现系统性下降。至关重要的是，这种崩溃在最容易诊断的类别中最为严重。在过敏性疾病和儿童风湿病病例中，对照模型的得分接近满分，而对抗式智能体的得分则大幅下降，约为 -6.0 分。这一现象证实了第4.2节提出的拒绝错误假说：当病例没有歧义时，引入批评者会造成了人为怀疑，进而导致裁决者过度思量证据，最终拒绝明确且正确的方案，转而选择辩论过程中生成的一个看似合理的干扰项。

4.4 局限性

尽管本研究对多智能体拓扑结构进行了全面评估，但仍须承认其中存在若干局限性。首先，我们依赖特定的大语言模型（GPT- 5.1），这会引入模型依赖性偏差。未来的研究应在不同模型系列（例如 Claude、Llama、Gemini）中复现这一设置。

其次，数据集仅包含源自 Orphanet数据库 的 302 个初级诊疗病例。由于这些病例专门聚焦于罕见病，研究发现未必能完全推广至常见的临床表现；在这类病例中，症状重叠更为频繁。此外，本研究将诊断过程视为静态的单轮推理任务。在真实临床情境中，诊断是一个迭代过程，涉及问诊、实验室检查和患者反馈 [22]。未来的研究应借助 DoctorAgent-RL [19] 和 MedAgentSim [23]等多轮对话框架，将拓扑结构评估拓展至动态环境中。

最后，我们还须考虑成本效益权衡。与对照模型相比，层级式和协作式模型需要消耗更多令牌，推理时间也显著更长。鉴于诊断准确率的提升幅度有限（<2%），复杂多智能体系统所带来的计算开销或许不适用于常规筛查，但在某些特定的高风险诊断子类别中可能仍有价值。

5. 结论

本研究系统评估了多智能体系统在临床推理中的表现，详细分析了架构设计如何影响诊断精度。我们的主要贡献如下：

- 我们针对 302 例罕见病病例，对四种不同的智能体架构——对照、层级式、对抗式和协作式——进行了实证比较。结果表明，**层级式**拓扑结构的表现最佳（准确率为 50.0%），协作式和对照紧随其后。

- 我们提出了一项新指标——推理差距（Δ = 推理召回率 - 诊断准确率），用于衡量知识缺失（检索失败）与判断失败（裁决失败）之间的差距。

- 我们揭示了对抗式拓扑中的一种严重失效模式，其推理差距高达 16.7 个百分点。与“辩论能够强化推理”这一预期相反，[8]，我们发现，强制采用唱反调者角色会引入人为怀疑，导致系统拒绝正确且概率很高的诊断。这种性能下降在“较容易”的诊断类别中最为严重，因为要求提出批评意见反而使明确的临床信号变得过于复杂。在较容易的诊断类别（如过敏性疾病、风湿性疾病）中，这种性能下降最为严重，因为辩论过程使明确的临床信号变得过于复杂。

- 我们进行了领域特定分析，结果表明，没有任何一种拓扑结构在所有情况下都占据优势。层级模型的总体诊断准确率最高，而协作方法在多器官病理方面展现出独特优势，例如呼吸系统疾病和泌尿生殖系统疾病；在这些情况下，整合不同诊断视角（放射学、内科学和病理学）至关重要。相反，对于心脏畸形这类高度模糊的类别，所有拓扑结构都同样难以应对，这表明在这些情况下，主要瓶颈在于病例数据不透明，而非智能体结构。

最终，这些发现表明，增加系统复杂性并不能保证推理能力更强。部署 AI 时，应超越静态架构，转向动态拓扑选择，由监督系统分配特定工作流程，例如：多器官病例采用协作拓扑，而常规病例则采用单智能体 [13]。

数据集可用性

本研究使用的数据集由Chen等人提供。 [9] 可通过以下链接访问：
https://github.com/geteff1/Multi-agent-conversation-for-disease-diagnosis

参考文献

[1] Tretow, I., Schwebel, M., Feuerriegel, S., Treffers, T., & Welpe, I. M. (2025). 大语言模型辅助对诊断准确率的影响：一项荟萃分析。载于 medRxiv。
https://doi.org/10.64898/2025.12.11.25341476

[2] Reese, J. T., Chimirri, L., Bridges, Y., Danis, D., Caufield, J. H., Gargano, M. A., Kroll, C., Schmeder, A., Liu, F., Wissink, K., McMurry, J. A., Graefe, A. S., Niyonkuru, E., Korn, D. R., Casiraghi, E., Valentini, G., Jacobsen, J. O., Haendel, M., Smedley, D., ... Robinson, P. N. (2025). 系统性基准测试表明，大语言模型尚未达到传统罕见病决策支持工具的诊断

准确率。https://doi.org/10.1101/2024.07.22.24310816

[3] Gupta, A., Sridhar, A. N., Mingole, B., & Yadav, A. (2025). 使用《豪斯医生》案例研究评估大语言模型在罕见病诊断中的表现。载于 arXiv [cs.CL]。https://doi.org/10.48550/arXiv.2511.10912

[4] Silveira, A. L., da Rosa Righi, R., & André da Costa, C. (2026). 用于临床决策支持的多智能体系统：系统性综述。应用软计算，188(114447)，114447。https://doi.org/10.1016/j.asoc.2025.114447

[5] Liu, P. R., Bansal, S., Dinh, J., Pawar, A., Satishkumar, R., Desai, S., Gupta, N., Wang, X., & Hu, S. (2025). MedChat：一种基于大语言模型的多模态诊断多智能体框架。2025 年 IEEE 第八届多媒体信息处理与检索国际会议（MIPR），456–462。

[6] Fan, Z., Wei, L., Tang, J., Chen, W., Siyuan, W., Wei, Z., & Huang, F.（2025年1月）。AI Hospital：在多智能体医疗交互模拟器中对大语言模型进行基准测试。载于第31届国际计算语言学会议论文集（第10183–10213页）。

[7] Feng, J., Zheng, Q., Wu, C., Zhao, Z., Zhang, Y., Wang, Y., & Xie, W.（2025）。M^3 builder：一种用于医学影像自动机器学习的多智能体系统。载于面向医学的智能体人工智能国际研讨会（第115–124页）。Cham：Springer Nature Switzerland。

[8] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I.（2023）。通过多智能体辩论提升语言模型的事实准确性与推理能力。载于 arXiv [cs.CL]。http://arxiv.org/abs/2305.14325

[9] Chen, X., Yi, H., You, M., Liu, W., Wang, L., Li, H., Zhang, X., Guo, Y., Fan, L., Chen, G., Lao, Q., Fu, W., Li, K., & Li, J. (2025). 通过多智能体对话式大语言模型增强诊断能力。npj数字医学，8(1)，159。https://doi.org/10.1038/s41746-025-01550-0

[10] Dinc, M. T., Bardak, A. E., Bahar, F., & Noronha, C. (2025). 大语言模型在临床诊断中的比较分析：常见及复杂医学病例中的性能评估。JAMIA Open，8(3)，ooaf055。https://doi.org/10.1093/jamiaopen/ooaf055

[11] Kim, Y., Jeong, H., Park, C., Park, E., Zhang, H., Liu, X., Lee, H., McDuff, D., Ghassemi, M., Breazeal, C., Tulebaev, S., & Park, H. W. (2025). 分层智能体监督：面向医疗安全的分层多智能体系统。载于 arXiv [cs.AI]. http://arxiv.org/abs/2506.12482

[12] Zuo, K., Jiang, Y., Mo, F., & Lio, P. (2024). KG4Diagnosis：面向医学诊断的知识图谱增强分层多智能体大语言模型框架。载于 arXiv [cs.AI].http://arxiv.org/abs/2412.16833
[13] Wang, Z., Wu, J., Cai, L., Low, C. H., Yang, X., Li, Q., & Jin, Y. (2025). MedAgent-Pro：通过推理智能体工作流实现循证多模态医学诊断。载于 arXiv[cs.AI]. http://arxiv.org/abs/2503.18968[14] Peng, Q., Cui, J., Xie, J., Cai, Y., & Li, Q. (2025). 推理树：通过基于证据树的多智能体推理实现复杂医学诊断。第33届ACM国际多媒体会议论文集, 1744–1753.

[15] Ji, X., Li, Z., Zeng, L., Hu, L., Wang, Y., Zhang, K., Shi, L., Wei, M., Chen, L., Guo, L.,Dong, J., Wang, A., Sun, L., Song, Y., Wang, H., Wang, J., Lei, Y., Yue, W., Zhao, Z., ... Li, K.(2026). 在基层医疗中利用多智能体系统对中轴型脊柱关节炎进行早期诊断。npj数字医学。 https://doi.org/10.1038/s41746-026-02372-4

[16] Dąbrowicki, W., Rusiecki, A., & Jeleń, Ł. (2025). Med-agent：一种用于多模态癌症诊断的混合人工智能智能体。Procedia Computer Science，270，3290–3299。 https://doi.org/10.1016/j.procs.2025.09.454

[17] Liu, Q.、Hu, Z.、Huang, T.、Niu, Y.、Zhang, X.、Ma, S.、Lin, C.、Huat, G. K.、Kwon, H. E.、Gao, F.、Sun, X.、Ying, Z. 和 Qiang, G.（2026）。EvoMDT：用于多癌种结构化临床决策的自进化多智能体系统。npj数字医学，9（1），124。https://doi.org/10.1038/s41746-025-02304-8

[18] Thakrar, K.、Basavatia, S. 和 Daftardar, A.（2025）。构建临床协作：用于多模态医学视觉问答的多智能体推理系统。载于 arXiv [cs.AI]。http://arxiv.org/abs/2507.05520

[19] Feng, Y.、Wang, J.、Zhou, L.、Lei, Z. 和 Li, Y.（2025）。DoctorAgent-RL：用于多轮临床对话的多智能体协作强化学习系统。载于 arXiv [cs.CL]。 https://doi.org/10.48550/arXiv.2505.19630

[20] Ai, R., Pan, Y., Simchi-Levi, D., Tambe, M., & Xu, H.（2025）。超越多数投票：利用高阶信息进行大语言模型聚合。载于 arXiv [cs.LG]。http://arxiv.org/abs/2510.01499

[21] Panickssery, A.、Bowman, S. 和 Feng, S.（2024）。大语言模型评估器能够识别并偏好自身的生成结果。《神经信息处理系统进展》，37，68772-68802。

[22] Wang, S.、Jiang, S.、Gao, Y.、Wang, B.、Gao, S. 和 Zhuang, X.（2025）。借助临床会诊流程赋能医疗多智能体，实现动态诊断。载于 arXiv [cs.AI]。
http://arxiv.org/abs/2503.16547

[23] Almansoori, M.、Kumar, K.与Cholakkal, H.（2026）。MedAgentSim：面向真实临床互动的自演化多智能体仿真。载于《计算机科学讲义》（第362–372页）。Springer Nature Switzerland。