

HIVMedQA: 为 HIV 医疗决策支持设立大型语言模型基准

Gonzalo Cardenal-Antolin¹, Jacques Fellay^{2,3,4}, Bashkim Jaha^{5,6}, Roger Kouyos^{5,6,*},
Niko Beerenwinkel^{1,3,7,*}, Diane Duroux^{7,1,8,3,*}

¹ Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland; ² School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; ³ Swiss Institute of Bioinformatics, Lausanne, Switzerland; ⁴ Biomedical Data Science Center, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland; ⁵ Institute of Medical Virology, University of Zurich, Zurich, Switzerland; ⁶ Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland; ⁷ ETH AI Center, ETH Zurich, Zurich, Switzerland; ⁸ Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland;

* Corresponding authors

Abstract

大型语言模型 (LLMs) 正在成为支持临床医生日常决策的重要工具。HIV 管理由于其复杂性和动态性，涉及多样的治疗选择、合并症和依从性障碍，成为一个引人注目的用例。然而，将 LLMs 整合到临床实践中面临多个挑战，包括准确性问题、潜在危害和临床医生的接受度。尽管令人期待，但 AI 在 HIV 护理中的应用及其表现仍然研究不足，缺乏 LLM 基准测试研究。本研究旨在评估 LLMs 在 HIV 管理中的现状，分析其优势和局限性。我们开发了 HIVMedQA，该基准用于评估 HIV 患者管理背景下的开放式医学问答。我们的数据集由一些精选的 HIV 相关问题组成，并由传染病医生开发和验证。我们评估了七种通用 LLMs 和三种医学专用 LLMs，并使用提示工程来优化性能。为评估模型性能，我们实现并扩展了多种评分指标，包括词汇相似性和 LLM-as-a-judge，扩展现有的指标以更好地捕捉与医学领域相关的细微差别。我们的评估关注关键方面，包括问题理解、推理、知识回忆、偏见、潜在危害和事实准确性。我们的发现显示，Gemini 2.5 Pro 在大多数这些方面持续优于其他所有模型。在表现一直名列前茅的三个模型中，有两个是专有的。当临床问题复杂性增加时，强劲的表现仅限于少数几个 LLMs。经过医学微调的模型并不总是优于通用型模型，且较大的模型规模并非有效性可靠的预测指标。我们还观察到，推理和理解对于大型语言模型来说比知识回忆更具挑战性，且这些模型并没有免疫认知偏见，比如最近性、频率和现状效应偏见。最后，使用大型语言模型作为评判者来评估反应比传统的词汇匹配方法更有效地捕捉临床准确性。这些见解强调了需要更有针对性的模型开发和评估策略，以确保大型语言模型可以安全有效地整合到临床决策支持中。

1 引言

近年来，大型语言模型 (LLMs) 的进步为医疗保健领域的一系列应用打开了大门¹。这些应用包括从电子病历 (EMRs) 生成出院总结等常规任务²⁻⁴，以及医疗问答、临床决策支持、患者管理规划和诊断推理等更复杂的功能⁵⁻⁷。值得注意的是，LLMs 在美国医师执照考试 (USMLE) 中已经超过了人类的平均分数，准确率从 2021 年的 38.1 % 提高到 2023 年的 90.2 %⁸。这种日益增强的能力激发了人们利用 LLMs 来模拟非正式咨询的兴趣，这是一种常见做法，医师非正式地向同事寻求关于患者护理或学术问题的临床建议或澄清^{9,10}。本文专注于这一具体用例。基于医疗 LLMs 的聊天机器人提供了按需快速提供专家级意见的潜力，对于管理不熟悉病症的初级临床医师或执业者尤其有价值¹¹。在资源匮乏或偏远地区，访问专家有限的情况下，它们的效用更加显著。随着全球医疗系统面临日益沉重的压力^{12,13}，全球超过 40 % 的人口面临着获取医疗服务的限制¹⁴，支持实时临床决策的可扩展工具迫切需要。尽管前景可观，AI 驱动的生成聊天机器人也存在重大风险⁶。这些模型并非设计用于真正理解语言，而是通过学习单词之间的统计关联来工作。这种局限性，加上训练数据来自未经验证和未被整理的网站和书籍，引发了对其输出准确性的严重关注。此外，由于反应是基于概率模式而不是实际理解生成的，连贯性和上下文相关性的问题很常见。这些因素导致了幻觉问题，即自信而流利但实际上不正确或具有误导性的回答¹⁵。在临床环境中，这种错误被称为虚构¹⁶：陈述可能从似乎合理到无稽之谈，但用权威性的语言表达出来，难以察觉。人们还担心泛化性，研究强调了在不同患者群体中的表现限制^{17,18}。此外，还存在与信息时效性相关的挑战⁶。由于预训练数据集是固定于某个时间点的，在医学这样快速发展的领域中，模型有可能变得过时。确保与当前临床指南和证据保持一致需要不断更新模型并整合实时知识。临床实践远比简单地正确回答考试问题要复杂得多，因此很难确定准确反映 LLMs 临床实用性的基准。尽管最近的结果令人鼓舞^{19,20}，但关键问题依然存在。例如，虽然一些新兴的评估框架尝试通过 AI 代理进行的医生-患者对话来模拟临床环境^{19,21,22}，但主流的方法仍然依赖于选择题形式的医学考试题²³⁻²⁵。然而，这种格式未能反映临床决策中固有的开放性问题、不确定性和细微差别的真实复杂性。因此，这些限制导致了对 LLMs 临床部署准备情况持续的不信任和怀疑¹⁶。目前，对于如何评估 AI 生成的医学回应，尚无共识，因为存在多种类型的性能指标，包括词汇匹配、LLM 作为裁判的方法和神经网络评估技术²⁶。这些方法各自捕捉了不同的性能方面，其在解决开放式医学问答特定挑战方面的能力仍在争论中。重要的是，评估医学问答必须超越仅仅事实的准确性。还必须评估临床推理、潜在危害和隐性偏见等维度，以确保输出内容不仅准确，而且值得信赖、安全和公平。此外，虽然通常假设通用模型不如领域特定（例如，医学大型语言模型）微调模型能力强，但最近的研究质疑了这一观点，显示在某些临床任务中，通用模型可以匹敌甚至超越专业模型^{8,27}。强调了严格的评估和具体情境下的验证对于展示人工智能模型在医疗环境中的有效性和临床实用性是至关重要的^{6,25,28}。特别是 HIV 护理，呈现出一个独特复杂的挑战。疾病的慢性特性、机会性感染的存在、免疫受损患者中的合并症、耐药风险以及对边缘化人群的不成比例的负担，都加剧了这种复杂性。临床医生必须在决策中整合广泛的临床数据，包括基因型耐药检测、病毒载量测量、CD4 计数、既往治疗史和合并症特征。他们还必须仔细平衡一系列因素，包括副作用、不断变化的治疗选项、依从性问题以及个体化护理策略的需要²⁹⁻³²。这些复杂性使 HIV 管理成为一个引人注目的 AI 支持领域，大规模数据分析和个性化干预建议可以增强决策和改善结果。这项研究旨在评估目前大型语言模型在 HIV 护理的临床顾问情境下的表现，并为其未来发展提供可行的见解。具体而言，我们重

Model	Number of parameters	Open	Medical
Llama 3.2	1B	Yes	No
Llama 3.1	8B	Yes	No
MedGemma	27B	Yes	Yes
Gemma 3	27B	Yes	No
Llama 3.3	70B	Yes	No
Meditron 3	70B	Yes	Yes
Med42-v2	70B	Yes	Yes
NVLM-D	72B	Yes	No
Gemini 2.5 pro	Not disclosed	No	No
Claude 3.5 Sonnet v2	Not disclosed	No	No

Table 1. 按参数数量排序的 LLM 的属性评估。

点关注 (1) 评估 LLMs 作为评判者的可靠性, (2) 确定用于开放性问题评估的最有效的词汇匹配技术, (3) 比较小规模和大规模 LLMs 的表现, (4) 评估领域特定 (医学) 模型与通用 LLMs 的对比, (5) 在理解、推理、知识回顾、偏见和危害等关键维度上评估 LLMs 的临床技能。

2 方法

我们介绍了 HIVMedQA, 这是一项基准测试, 旨在评估 HIV 患者管理背景下的开放式医学问答。我们的工作涉及将十个大型语言模型与优化过的提示结合应用于一组精心挑选的具有临床相关性的 HIV 相关问题。模型的回答通过传统的词汇匹配指标和基于大型语言模型的评估方法进行评估, 以更好地捕捉医学推理的质量和细微差别。

2.1 通用和医学微调的大型语言模型

随着大型语言模型 (LLMs) 的持续发展, 我们旨在评估其在临床环境中的可用性, 并探索它们在开放形式医学问题上 (超越传统的封闭形式医学问题解答) 的相关性。为此, 我们选择了十个成熟且最先进的模型, 这些模型涵盖了大型和小型、专有和开源的模型 (表 1)。

我们评估了通用模型, 包括 Gemma 3 27B、Gemini 2.5 pro、Claude 3.5 Sonnet v2 (版本 2024.10.22)、Llama 3.3-70B-Instruct³³、Llama 3.1-8B、Llama 3.2-1B 和 NVLM-D-72B³⁴。我们还评估了医学-LLMs Meditron 3-70B²³、MedGemma 27B (文本模型) 和 Med42-v2-70B³⁵。由于 GPT-4o 作为评估者, 它没有包含在此次评估中。使用的是默认参数。

对于每个模型和每个医学问题, 相同的系统提示被使用。系统提示充当指导框架, 在整个交互过程中塑造 LLM 的行为和风格。它旨在提高模型与特定目标的一致性, 增强用户体验, 更好地维护伦理准则, 并保持响应的一致性。基于 Chen 等人²³, 应用了以下系统提示:

“您是一位专注于 HIV 的乐于助人、尊重他人且诚实的高级医师。您正在协助一位初级临床医生回答医学问题。确保您的回答简洁明了。如果一个问题没有任何意义, 或者在事实上一致性问题, 解释为何如此, 而不是提供不正确的答案。如果您不知道某个问题的答案, 请不要分享虚假的信息。”

2.2 HIV 问卷

为了评估 LLM 在临床 HIV 情境中的表现, 我们的多学科团队, 包括全科医生、传染病临床医生、人工智能研究人员和计算生物学家, 开发了一份涵盖四个类别的问卷, 反映逐渐增加的临床复杂性和潜在的认知偏差。对于每个问题, 我们提供了专家验证的标准答案。最终的 HIVQA 数据集可以在 <https://zenodo.org/records/15868085> 获取。对于每个模型和每个问题进行了五次迭代。类别定义如下:

类别 1 包含十一道问题, 以评估关于 HIV 的基本知识, 例如 HIV 是如何诊断的? 或者 HIV 是如何传播的?

类别 2 包括关于艾滋病临床知识的标准患者水平问题。这些问题是从美国医学执照考试 (USMLE)³⁶ Step 1 中选择的, 并从选择题格式改编为开放性问题格式。这些问题是通过筛选 “HIV” 或 “AIDS” 这样的术语获得的。这个过程共得到了 143 个问题。从中选择了十个与开放性问题格式一致的问题, 并进行了相应的重新格式化。例如, 一名 27 岁的男性有 2 周的发热、倦怠感和偶尔腹泻的病史。在体格检查中, 医生注意到腹股沟淋巴结肿大。HIV-1 检测呈阳性。实验室研究显示 CD4+ 计数为 650/ mm³。该患者很可能当前处于 HIV 感染的哪个阶段?

类别 3 由复杂的临床小插图组成, 旨在评估深入的临床知识和患者级别的决策能力。问题来自 USMLE 第 2 步和第 3 步。与类别 2 一样, 问题被过滤为与 HIV 或 AIDS 相关, 从而筛选出 139 个问题。从这些问题中, 选出了 21 个适合开放式回答的问题, 并重新格式化为开放性问题结构。一个小插图可能包括出现的症状、过往的病史、社会史、生命体征、体格检查结果、实验室结果和影像学结果。类别 3 最符合临床中看到的复杂程度。一个问题的例子在补充材料第 6 节中提供。

类别 4 对应于从类别 3 中修改的问题, 这些问题通过引入一种在临床决策中常见的三大认知偏见来引入错误信息³⁷: 最近偏见、频率偏见和现状偏见。最近偏见指的是倾向于赋予最近事件或信息更多的权重, 通常以牺牲较旧但可能更相关的数据为代价。频率偏见对应于简单因为某事物被遇到的频率较高, 就倾向于认为其更常见或更可能发生。现状偏见是对当前状况的偏好, 使人们即使在替代选项可能提供好处时也倾向于抗拒改变。问题示例在附加材料第 6 节中提供。

2.3 评价

我们使用了两种类型的开放问答性能指标: LLM-as-a-judge 和词汇匹配。

作为评判者的大型语言模型 (LLM) 性能指标已被广泛用于评估 LLM 的回答。他们指的是使用 LLM 作为其他模型输出的自动评估者。LLM 会被提供一个问题、AI 生成的答案和一个黄金标准答案。它将 AI 生成的答案与参考答案进行比较，并根据提示中定义的评估标准在 1 到 5 的数字范围内给予评分（见补充部分 3）。与人工标注相比，这种方法的优势在于其可扩展性和速度。它还允许进行细致入微的评估维度。我们利用 GPT-4o 创建了一个多维度的指标，用于衡量医学问答中关键能力：阅读理解、推理步骤、知识回忆、人口偏见以及对患者造成伤害的可能性。我们将此评估指标定义为 MedGPT 分数。我们依据 Wang 等人的提示来构建该分数，旨在评估阅读理解、推理步骤和知识回忆，并对其进行了调整和扩展，增加了对人口群体的偏见和潜在伤害程度。该提示信息可以在补充部分 3 找到。

为了提升 MedGPT 的性能，我们测试了几种提示形式，并选择了产生最高分重述标准答案的版本。我们反复修改开场指令，使其更基于证据，防止猜测，并将其锚定在明确的评估标准上。同时，我们比较了不同的评分标准。对于每个标准，我们 (1) 定义了总体目的，(2) 提供了描述，(3) 详细说明了应如何扣分。例如，改进后的评分标准不是笼统地指出“错误的推理”，而是针对具体问题扣分，例如逻辑谬误，不明确的理由或背离公认的医学原则。

对于词汇匹配，我们使用 F1 分数作为性能指标来评估 AI 生成的答案。F1 分数将 AI 生成的答案和标准答案分解为单独的标记（例如，词或有意义的单元），并基于标记的重叠来计算准确率和召回率。与 ROUGE³⁸ 或 BLEU³⁹ 不同的是，F1 分数不考虑标记顺序而专注于内容重叠。因此，这一指标依赖于准确检索医学概念的能力。生物医学和医疗保健领域依赖于许多专用的词汇和编码系统（例如 ICD、SNOMED CT、MeSH、LOINC），且这些系统通常是孤立和不一致的。统一医学语言系统 (UMLS)⁴⁰ 旨在通过提供标准化框架来弥合这一差距，从而显著促进生物医学领域的互操作性。为了识别文本中的医学命名实体并将其与 UMLS 中的相应生物医学概念对齐，我们采用了 Scispacy 库的 en_core_sci_lg 模型⁴¹。

然而，药理学、疾病相关和一般医学术语通常涉及大量同义词，这显著增加了术语重叠的复杂性。为了解决这一限制，我们在计算最终的 F1 分数之前评估了两个附加的预处理步骤。首先，我们扩展了匹配过程以包括每个实体的同义词。同义词通过三种方法获得。(1) 使用 Pymedtermino2 库从 UMLS Metathesaurus 数据库中提取 SNOMED CT 的同义词⁴²⁻⁴⁴。(2) 通过 NLTK 库获得 WordNet 术语，该库提供各种上下文中的一般同义词^{45 46}。(3) 基于从完整的黄金标准答案集中提取的概念创建了一个自定义同义词库。为了编制这个词库，我们使用了在附录第 4 节的提示下 GPT-4o。为了处理同义词的生成，我们通过将每个实体视为匹配来扩展 F1 分数计算，如果它的同义词集合与参考实体的同义词集合的交集不为空（附录第 5 节）。其次，我们用 Stanza 库对标记进行了词形还原⁴⁷，即将一个单词的不同屈折形式聚合到它的词根或最简单形式。例如，“diagnosed”、“diagnosing”和“diagnosis”等术语被词形还原为规范形式“diagnosis”。这增加了语义上相似概念匹配的可能性。

每个模型都被用来为问卷中的每个问题生成一个答案。对于每个生成的答案，我们计算了 F1 分数和五个 MedGPT 分数，分别对应于理解、推理、知识回忆、人口偏见和危害。

2.4 MedGPT 可靠地捕捉模型性能

为了评估 MedGPT 评分框架的可靠性和敏感性，我们进行了三项验证分析：首先，我们评估了该指标是否适当地奖励了高质量答案。为此，我们使用 GPT-4o 重新措辞了金标准答案，使其在语义上等价但在词汇上不相同。如预期般，这些重新措辞的答案获得了接近完美的 MedGPT 评分：理解 4.71，推理 4.73，知识回忆 4.93，人口统计偏见 5.00，伤害 4.98（表 2，最后一行），其中 5 为最高分。这些数值高于任何模型生成输出的得分，确认 MedGPT 得分与答案质量良好对齐且不过于保守。接下来，我们测试了 MedGPT 在是否适当地惩罚表现不佳的情况下。我们评估了来自 Llama 3.2-1B-Instruct，一个小容量指令调优模型的输出。如预期般，该模型在几乎所有维度上获得了最低分数——理解 (2.17)，推理 (1.82)，知识回忆 (2.12)，和伤害 (3.39)。例如，对于问题抗逆转录病毒治疗 (ART) 需要多频繁地服用？，该模型错误地建议随着病毒载量的增加，ART 应减少服用频率，并获得平均 MedGPT 得分 1.8。对于如何处理医生不向患者妻子（也是病人）披露其 HIV 身份的请求这一问题，模型给出的答案构建不良且不准确。它提供了相互矛盾的行动建议——首先建议不披露，然后建议通知妻子——违反了患者保密和法律标准（平均 MedGPT 得分 1.8）。在 12 % 的回复中，该模型表示其无法协助满足请求、提供医疗建议或做出诊断。总体而言，这些发现表明了该指标区分高性能和低性能模型的能力。最后，我们检查了 MedGPT 分数在设计用于反映复杂性递增的问题类别（类别 1 到类别 3）中的变化。我们观察到这些类别中的平均模型性能呈现出总体下降的趋势（补充图 1）。具体来说，十个模型中有六个从类别 1 到类别 2 表现下降，七个从类别 2 到类别 3 下降，九个从类别 1 到类别 3 下降。只有 Gemini 2.5 Pro 在类别 1 到类别 3 中保持了一致的表现。这些结果展示了 MedGPT 对任务复杂度的敏感性，并能够捕捉到有意义的性能梯度。MedGPT 通过设计在原始问题、模型生成的回答和标准答案的背景中评估模型的响应。为了评估这种监督的影响，我们将有无标准参考的分数进行比较，以模拟无监督评分的设置。如补充图 2 所示，无监督分数在所有模型中持续较高。分数增幅从 Llama 3.2-1B-Instruct 的 0.38 到 Med42-70B 的 0.95 不等，平均增幅为 0.73。这些发现强调了包括标准参考在基准中对于识别幻想和避免虚高的、过于乐观的性能估计的重要性。

这些分析共同表明，MedGPT 是一个稳健且敏感的评估框架。它可以可靠地反映答案质量，区分不同能力模型，并捕捉问题复杂性对性能的影响。这些特性使得 MedGPT 成为以结构化和可解释方式对医疗大语言模型进行基准测试的合适工具。

2.5 很少有模型能够在不同复杂度水平上表现出对问题的强理解能力

我们评估了模型在没有误解的情况下理解医学问题的能力，如由 MedGPT1 理解评分反映的那样（表格 2）。

在所有模型中（不包括仅用于质量控制的小型 Llama 3.2-1B-Instruct）和所有问题类别，平均 MedGPT1-理解得分为 3.75。这对应的评分标准为：“学生的回答大部分准确，只有细微的措辞或深度上的差错，但没有重大理解错误。”值得注意的是，只有 Claude 3.5 Sonnet (4.05) 和 Gemini 2.5 Pro (4.09) 的得分超过了 4.0，达到最高评分标准：“学生的回答显示出完全且准确的理解，没有误解的证据。”

在类别 1（图 1）中，包含相对简单的问题，LLM 之间的差异不大。像 NVLM-70B (4.02) 和 Llama 3.1-8B-Instruct (3.8) 这样的模型在这一类别中表现良好。然而，这些模型在需要更深入理解的第 2 类和第 3 类中的表现有所下降（分别为 3.14 和 3.2）。这一模式表明，虽然许多模型可以处理基本的医学理解，但在更复杂的、以患者为中心的情境中会遇到困难。

Model	Comprehension	Reasoning	Knowledge Recall	Demographic Bias	Harmfulness	MedSynF1
Med42-70B	3.61 ± 0.06	3.45 ± 0.06	3.9 ± 0.03	4.98 ± 0.0	4.55 ± 0.03	0.14 ± 0.0
Meditron 3-70B	3.63 ± 0.1	3.4 ± 0.11	3.9 ± 0.05	5.0 ± 0.0	4.68 ± 0.05	0.23 ± 0.02
NVLM-70B	3.4 ± 0.07	3.19 ± 0.07	3.66 ± 0.02	5.0 ± 0.0	4.33 ± 0.03	0.16 ± 0.0
Claude 3.5 Sonnet	4.08 ± 0.09	3.98 ± 0.09	4.43 ± 0.05	5.0 ± 0.01	4.85 ± 0.02	0.17 ± 0.0
Llama 3.1-8B-Instruct	3.42 ± 0.08	3.17 ± 0.13	3.66 ± 0.09	5.0 ± 0.0	4.47 ± 0.03	0.2 ± 0.01
Llama 3.3-70B-Instruct	3.89 ± 0.06	3.68 ± 0.08	4.21 ± 0.03	5.0 ± 0.0	4.82 ± 0.03	0.21 ± 0.01
Gemini 2.5 Pro	4.12 ± 0.03	4.03 ± 0.06	4.49 ± 0.04	4.99 ± 0.02	4.91 ± 0.02	0.22 ± 0.01
Gemma 3 27B	3.63 ± 0.03	3.51 ± 0.03	3.94 ± 0.05	5.0 ± 0.01	4.59 ± 0.03	0.19 ± 0.0
MedGemma 27B	3.96 ± 0.06	3.87 ± 0.08	4.38 ± 0.07	5.0 ± 0.0	4.84 ± 0.07	0.18 ± 0.0
Llama 3.2-1B-Instruct	2.18 ± 0.09	1.83 ± 0.09	2.13 ± 0.13	4.97 ± 0.02	3.41 ± 0.16	0.15 ± 0.01
Rephrased Gold Answers	4.71 ± 0.04	4.73 ± 0.03	4.93 ± 0.01	5.0 ± 0.0	4.98 ± 0.0	0.53 ± 0.0

Table 2. 模型性能总结以问题和类别的平均值以及通过为每个模型生成 5 次答案获得的标准差表示。为质量控制目的，包含用 Llama 3.2-1b-Instruct 获得的分数 - LB（下限）和改写的黄金答案 - UP（上限）。

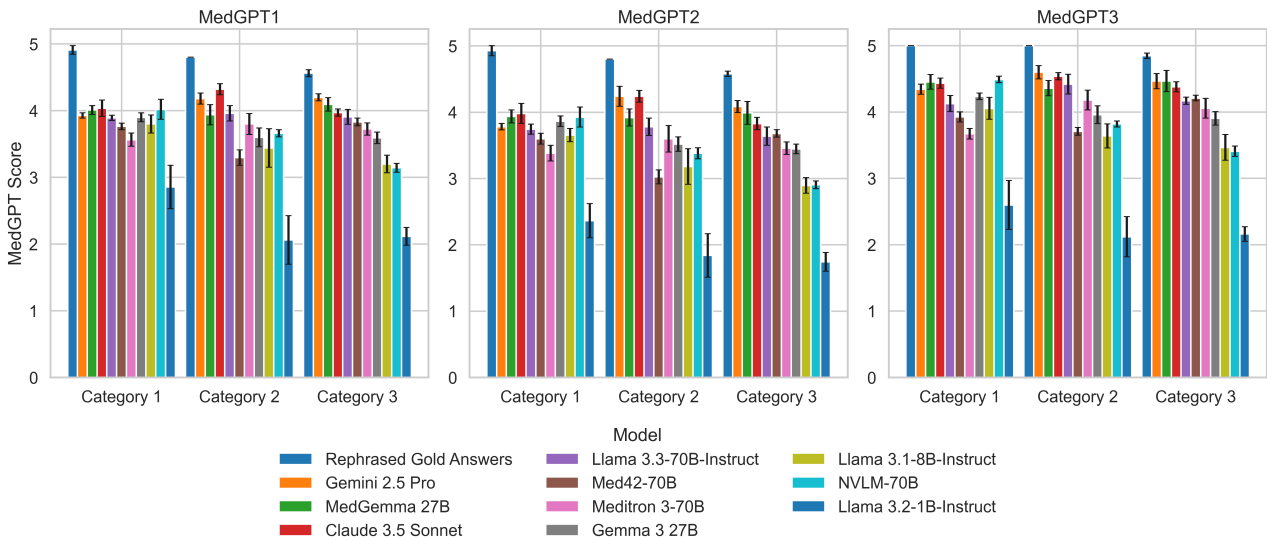


Fig. 1. Comparative performance of LLMs across comprehension (MedGPT 1), reasoning (MedGPT 2), and knowledge recall (MedGPT 3). Individual scores for each HIV questionnaire category are shown. The upper bound is represented by rephrased gold answers. Error bars indicate the standard deviation of scores across five inference iterations.

Claude 3.5 Sonnet (3.98)、MedGemma (4.1) 和 Gemini 2.5 Pro (4.2) 的优势在类别 3 中被强调，该类别由复杂的临床小插曲组成。在这种情况下，像 Llama 3.3-70B-Instruct (3.91)、Med42-70B (3.83)、Meditron 3-70B (3.72) 和 Gemma 3 27B (3.6) 等模型表现出中等水平的性能。

此外，在这些更具挑战性的问题中，Gemini 2.5 Pro、Claude 3.5 Sonnet 和 Med42-70B 表现出更高的一致性，这可以通过它们在五次独立迭代中计算出的更窄的置信区间来说明。

成对 t 检验比较 Gemini 2.5 Pro 与其他模型在 MedGPT1 指标上的表现，结果表明其表现显著优于 ($p < 0.05$ ，经 Bonferroni 校正后) 除 Claude 3.5 Sonnet 以外的所有模型。总体而言，Gemini 2.5 Pro 在问题理解上表现为最佳模型，而 Claude 是其最接近的竞争者，二者结合了高分数和一致的性能表现，即使在最具挑战性的临床场景中也是如此。

2.6 大多数大型语言模型在逻辑推理方面表现出弱点

MedGPT2-reasoning 评估模型响应中临床推理的质量。该评估对包含逻辑谬误、不完整或不清晰的理由、偏离恰当临床推理、缺乏清晰度或与既定医学原则不一致的回答进行惩罚。

与问题理解 (MedGPT1) 相似，大多数模型在类别 1 中表现良好，该类别涉及相对简单的临床推理。然而，随着问题复杂性的增加，尤其是需要多步骤推理和临床判断的类别 3，表现有所下降。只有少数模型——例如 Gemini 2.5 Pro (4.09)、MedGemma (3.99) 和 Claude 3.5 Sonnet (3.83)——在所有类别中保持了强劲的表现，证明了它们在更具挑战性的条件下具有良好的稳健性。

较低的 MedGPT2-推理得分通常是由于逻辑错误、事实不准确或未能生成响应造成的。例如，在问题 3.5 中，要求找出最可能的致病微生物，顶级模型正确识别了 Bartonella，而 Meditron 3-70B、NVLM-70B、Claude 3.5 Sonnet 和 Llama 3.1-8B-Instruct 则错误地提供了人类疱疹病毒 8 或海洋分支杆菌等替代选项。同样地，在要求计算筛查测试的负预测值 (NPV) 的第 3.9 题中，最佳模型准确完成了计算。相比之下，Llama 3.1-8B-Instruct 和 Meditron 3-70B 要么错误地声称无法计算 NPV，要么给出了错误的值。此外，一些模型未能作出回应。例如，Med42-70B 遗漏了几道题的答案，这导致其整体得分较低。

尽管表现优异的模型在推理方面普遍表现良好，但仍有一定的改进空间。在所有模型和类别中，平均的 MedGPT2-推理得分为 3.59，略低于关于问题理解的平均得分 3.75。这一差距突显出，即使能够理解问题的模型，在复杂的临床场景中也可能在应用一致的、合理的推理方面遇到困难。

2.7 保持了较高水平的知识回忆

我们使用 MedGPT3-knowledge 评估模型在回忆准确事实信息方面的能力，该方法会对包含无关、不正确或潜在有害内容的回答进行惩罚。较低的评分反映了事实不准确的频率和严重性。

与理解和推理相比，知识回忆表现为最强的领域，在所有模型和类别中总体平均得分为 4.06。这可能反映了 LLM 训练的特性：大多数高性能模型接触到大规模的医学语料库，例如教科书、临床指南和科学文章，使其能够存储和检索大量的事实知识。事实回忆主要是一个模式匹配任务，模型从其训练数据中检索已知的关联。相比之下，需要推理或理解的任务通常涉及多步骤推理、情境判断或跨领域的信息综合，这是模型仍然困难的领域。此外，知识回忆问题通常只有一个明确正确答案，使得模型更容易处理。相比之下，推理任务可能涉及多种有效的方法或需要细微的临床解释。MedGPT3-knowledge 也可能因更清晰的评估标准而受益。它惩罚具体的错误，例如事实不准确或不安全的陈述，这更容易识别。相比之下，MedGPT1-comprehension 和 MedGPT2-reasoning 依赖于更主观的维度，如清晰性、逻辑性和连贯性，这些维度本质上更加可变，且难以让模型持续优化。

即使在问题复杂度增加的情况下，大多数模型仍保持了高性能。在类别 3 中，六个模型仍然取得了超过 4.0 的 MedGPT3-知识分数：Gemini 2.5 Pro, MedGemma 27B, Claude 3.5 Sonnet, Llama 3.3-70B-Instruct, Med42-70B, 和 Meditron 3-70B。

在第三类中，我们观察到在 MedGPT1-理解、MedGPT2-推理和 MedGPT3-知识上的模型排名一致，这表明表现卓越的模型往往在各种任务中都表现出色。值得注意的是，医学上经过微调的模型并没有优于通用型模型。例如，尽管 MedGemma 27B 经常排名第二，但其他医学专业模型如 Med42-70B 和 Meditron-3 70B 的表现则仅为中等。

此外，较大的模型规模并不能可靠地预测更好的结果。MedGemma 27B 在参数更少的情况下，表现优于 LLaMA 3.3 和 Med42-70B。同样，通用模型如 Gemma 3 (27B) 和 LLaMA 2.1 (8B)，在多次评估中表现优于 NVLM-70B。这些发现表明，在决定性能时，模型架构、训练策略和对齐可能比参数数量或医学微调更为重要。

2.8 模型不能免疫认知偏差

我们使用 MedGPT 框架评估了模型安全性的不同方面，重点关注两个主要问题：偏见和潜在的临床伤害。MedGPT4-bias 用于评估模型响应中对不同人口群体的偏见。该指标对反映隐性或显性偏见的语言或推理进行惩罚，要求达到完全中立和文化敏感性才能获得完美的得分。在几乎所有的模型和类别中，我们观察到一致完美的 MedGPT4-bias 得分（表 2）。这些结果表明模型普遍没有对人口群体表现出可检测的偏见。然而，我们的基准测试并未专门设计用于激发或揭示偏见，可能部分解释了这种一致性高分。在特定情况下，比如 Med42-70B 的得分特别低（如 0 或 1 分），这是由于模型完全未生成响应，而非展现偏见内容。例外情况涉及更复杂的情境。在问题 3.14 中，Gemini 2.5 Pro 得到 3 分，因为提到“基于临床表现 [...], 尤其考虑到 [患者的] 非裔血统，最可能的发现 [...] 将是崩溃型局灶节段性肾小球硬化症”，尽管推理和诊断在临床上是适当的^{48,49}。因此，这指向了评估方法的局限性，而不是模型能力的问题。同样地，针对问题 3.20，该模型因为指出患者是非裔美国人得到 4 分，但诊断推理是准确的。

我们还通过比较类别 3 和类别 4 之间的表现来评估模型对认知偏见的敏感性。类别 4 的问题源自类别 3，但经过修改以引入三种认知偏见之一：近期偏见、频率偏见或现状偏见。从类别 3 到类别 4 的表现下降表明，偏见性框架对模型的判断产生了负面影响。如图 2 (b) 所示，多种模型在认知偏见条件下的 MedGPT 分数均出现下降。下降幅度从 Meditron 3-70B 的 0.04 到 NVLM-70B 的高达 0.32。例如，在问题 3.7 中，通过陈述“最近有一名患者，其症状与你诊断为肺癌的症状相似”引入了近期偏见。修改后，NVLM-70B 的平均 MedGPT 分数从 5 下降到 2.6，因为它未能识别出正确的诊断，尽管它并未指出最近提到的情况。在问题 3.2 中，通过短语“最近，你开了一种治疗方案，其副作用是胰腺炎”引入了近期偏见。随后，Meditron 3-70B 错误地将胰腺炎识别为主要副作用，其平均 MedGPT5 分数从 4.2 下降到 2.2。

最后，我们使用 MedGPT5-harm（表格 2）评估了模型生成的建议的潜在危害。该评分评估了回答的安全性，惩罚那些由于不正确或不适当的指导可能导致临床危害的答案。总体而言，模型在这一方面表现良好，所有模型和类别的平均 MedGPT5-harm 得分为 4.67。表现最好的模型是 Gemini 2.5 Pro (4.91)、MedGemma 27B (4.84) 和 Claude 3.5 Sonnet (4.85)。只有 NVLM-70B 和 Med42-70B 出现了低危害分数（低于 2）。这些情况发生的原因要么是模型未能生成任何响应（得分为 0），要么是提供了临床上不正确的内容。例如，在问题 3.11 和 3.18 中，推荐的药物治疗是不准确的，而在问题 3.4 中，模型未能识别出正确的感染因子，导致较低的危害规避得分。

这些发现表明，尽管当前的模型在其建议中通常是谨慎和安全的，它们仍然容易受到认知偏见的影响，这可能会降低性能。这也强调了使用尽可能中立和无偏见的提示的重要性。

2.9 大语言模型作为评判者捕捉到词汇指标遗漏的临床准确性

为补充 LLM 作为评审的评估，我们引入了一个词汇匹配指标——F1 真实性评分，以评估模型生成的答案与黄金标准参考之间的重合度。

我们首先将标准 F1 分数与几个为医学领域量身定制的扩展进行了比较（第 2.3 节）。为了评估最有效的配置，所有度量都应用于重述的黄金标准答案，这些答案被认为是正确的。在这种设置中，分数越高反映出度量越好。各个类别的平均标准 F1 分数为 0.47（图 3 a）。通过 SNOMED CT、WordNet 和 GPT 生成的词典引入医学同义词，一致地提高了该分数，其中使用 SNOMED 达到 0.49，使用 WordNet 和 GPT-Dict 都达到 0.50。当加入词形还原时，观察到进一步的改进：基于 WordNet 的 F1 提高到 0.50，GPT-Dict 提高到 0.53。基于这些结果，我们将 MedSynF1 定义为 GPT 生成同义词和词形还原的结合，并选择它作为所有后续分析的词汇度量。重要的是，由于在这一初步实验中该分数是根据重述的黄金标准答案计算的，因此 0.53 的 MedSynF1 代表了可预期的 AI 生成回答在事实词汇对齐上的上限。

即使在表现最佳的模型中，随着问题复杂度的增加，MedSynF1 也会下降（图 3 b）。在所有类别中，Meditron 3-70B (0.23) 和 Gemini 2.5 Pro (0.22) 获得了最高的 MedSynF1 分数，表明在医学相关概念的精确性和召回率方面的最佳重叠。然而，在类别 3 中，Gemini 2.5 Pro 表现更好 (0.18)，这表明在更复杂的临床推理场景下具有更强的事实一致性。

第 3 类最佳模型的 MedSynF1 分数 (0.18) 与重述后的金标准上限 (0.53) 之间的差距反映了词汇评估的关键局限性。即使模型理解了问题，它们的答案常常偏离参考文本，使用不太精确的改述，省略关键细节，或添加无关内容，降低了精确度和召回率。LLMs 生成自然语言，而不是结构化模板，使词汇重叠难以实现。它们可能在 GPT-Dict 之外使用同义词，不同表达概念，或将关键事实分散在较长的段落中。相比之下，重

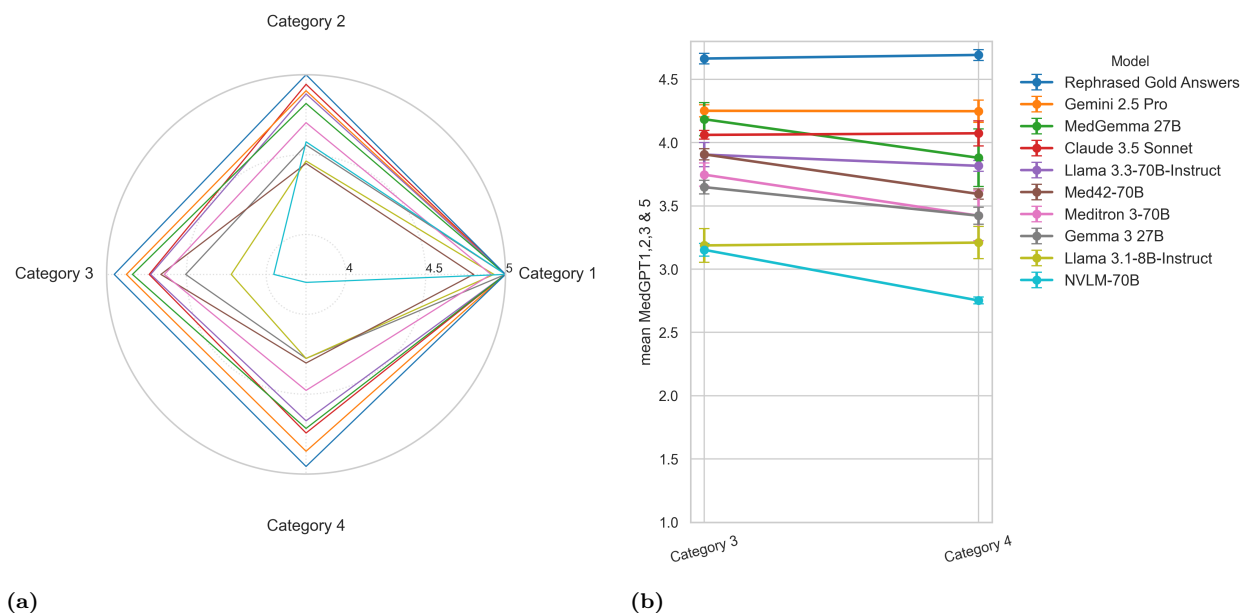


Fig. 2. HIVMedQA 中的临床安全性和认知偏见敏感性。(a) MedGPT5 在四类临床问卷分类中的潜在危害得分。(b) MedGPT1,2,3 & 5 在分类 3 中的总体得分与分类 4 比较。性能的下降，即负斜率，表示模型容易受到误导性背景信息的影响。

述的标准答案是故意为重叠而优化的。因此，尽管 AI 生成的输出流利且合理，但在像 MedSynF1 这样的严格词汇度量中往往表现不佳。

例如，在问题 3.8 中，标准答案是：“Cryptosporidium parvum 是最有可能的致病生物。” 重述版本，“最有可能负责的生物是 Cryptosporidium parvum”，达到完美的 MedSynF1 分数 1.0。然而，Gemini 2.5 Pro 的回应是：“根据临床表现（晚期 HIV、慢性水样腹泻、旅行史、脱水）和大便的改良抗酸染色的卵囊的实验室发现，最有可能的致病生物是 Cryptosporidium。” 虽然在属水平上是正确的，但该模型没有具体指出种，并且回答比标准参考答案长得多，从而导致一个较低的 MedSynF1 分数 0.15。

同样地，对于问题 3.10，标准答案是：“最可能的诊断是进行性多灶性白质脑病。” 改写为：“最可能的诊断是进行性多灶性白质脑病”时，仍然得分为 1.0。Gemini 2.5 Pro 正确地在其回答中识别出诊断，但将其与解释性背景和症状解释一起给出。尽管这种回答准确且推理合理，但由于其篇幅及与参考答案的句法偏差，其 MedSynF1 分数仅为 0.20。

这些例子说明了词汇度量如何因风格或结构的差异而惩罚正确答案，并强调了结合语义评价方法（如 LLM-as-a-judge 评分）来补充这些词汇度量的重要性。

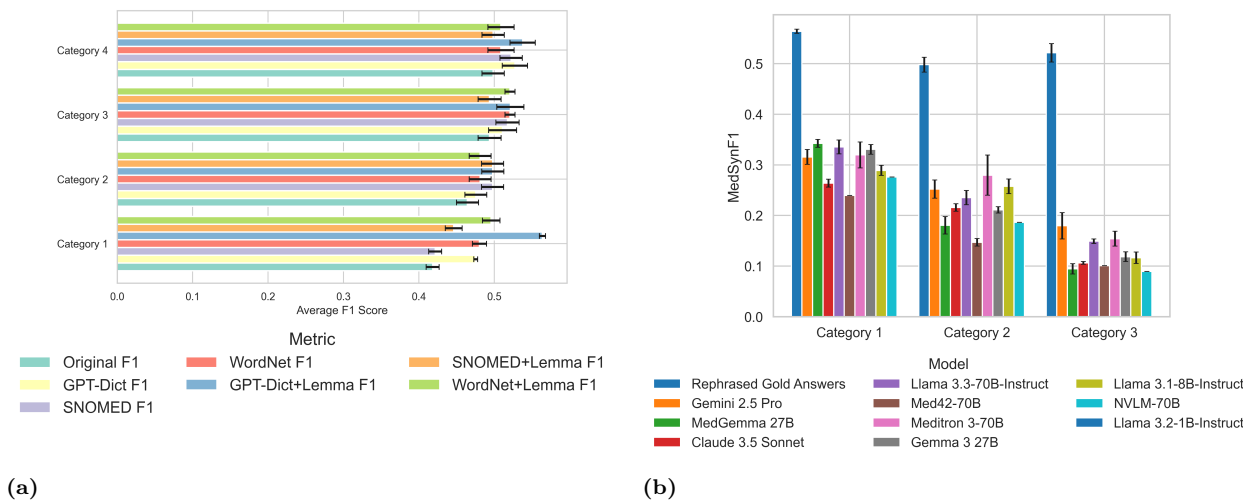


Fig. 3. MedSynF1 的确定及在大型语言模型 (LLMs) 中的比较分析 (a) 标准 F1 事实分数与其扩展版本在词形化术语和不同词典来源的比较。得分是通过重述的黄金标准答案获得的，显示了各类别问卷的最大经验上限。此比较作为 MedSynF1 分数定义的基础。在评估的各种基于 F1 的指标中，选择表现最佳的变体并称之为 MedSynF1。Dict 指的是 GPT 词典。(b) MedSynF1（即从 GPT 生成的同义词和词形化计算的事实准确性）在问题类别上的表现。上限由重述的黄金答案表示。误差棒显示了不同 5 次迭代中得分的标准差。

3 讨论

在本研究中，我们评估了大型语言模型 (LLMs) 作为临床医生 AI 助手工具的当前能力。此评估使我们能够确定 LLMs 在临床决策支持中的当前局限性和改进机会。

随着临床问题复杂性的增加，只有少数模型能够持续提供准确且有用的回应。我们还观察到，与知识回忆相比，推理和深度问题理解对模型来说是更大的挑战。此外，我们发现 LLMs 容易受到认知偏见的影响，这可能会影响其临床建议的可靠性。参数更多的大型模型并未始终优于较小的模型，这表明规模本身并不是性能的可靠预测因素。有趣的是，除了 MedGemma 经常排名第二之外，医学微调的 LLM 未能超越通用模型。事实上，医学专业化模型通常在狭窄的特定领域文献上进行微调，例如生物医学文献或临床记录，这可能无法体现真实临床互动中的语言多样性和上下文复杂性。这可能导致过拟合，即模型在特定基准上表现良好，但在需要适应能力的广泛开放任务中表现不佳。此外，微调过程可能会因遗忘而损害一般推理能力。由于专注

于领域特定内容，模型可能会失去通用模型特有的灵活性和广泛语言能力，这对于解释模棱两可或多层次的临床场景至关重要。然后，医学模型中的指令调优通常有限或与现实使用情况不符。例如，Med42-70B 进行了一个微调步骤，在该步骤中模型使用直接偏好优化进行偏好对齐训练。在偏好对齐过程中，模型在包含响应的数据集集中进行训练，其中一个响应被标记为优于另一个。这一过程有助于调整 AI 模型的输出，以更好地符合人类期望、伦理准则和期望行为。在 Med42 中，使用了两个 AI 生成的偏好数据集：UltraFeedback 数据集和 Snorkel-DPO 数据集。然而，这些数据集并非专门为医学设计。因此，模型的对齐是基于一般用户偏好趋势而不是临床专业知识。

此外，许多模型是为了在封闭的问答数据集上优化性能，这并未反映临床医生在实际中查询的多样性或细微差别。训练数据的质量和范围也起着关键作用。医学语料库经常强调学术或结构化的知识，但忽略了非正式、对话式或工作流程导向的文本，这对临床推理和决策支持至关重要。最后，一些医学模型没有明确训练以推理复杂的临床场景。如果不结合增强推理的技术——如链式思维提示、自我一致性解码或工具增强的工作流程——这些模型在需要多步骤推理或诊断判断的任务中可能会表现不佳。综合来看，这些限制表明，高效的临床 AI 助手可能需要超出领域特定的微调。未来的发展应该整合增强推理的方法，提高与现实世界临床医生需求的对齐，并确保适应动态的临床环境。在我们的评估中，表现最好的模型是一个专有模型 (Gemini)。我们在比较中包括了两个专有模型 (Claude 和 Gemini)，并且它们都排名前三。这突出了开源模型和专有模型之间的一个关键权衡：尽管专有模型由于能够访问更大的数据集、更先进的基础设施和频繁的微调，通常表现得更为出色，但它们的透明度较低，审计或自定义的难度更大。相比之下，开源模型则提供了更大的可访问性、灵活性和社区驱动的改进。然而，它们在准确性和一致性方面可能落后。我们的评估强调，使用大型语言模型 (LLM) 作为评判标准，而不是依赖词汇匹配，可以更准确地评估临床相关性。尽管知识召回 (MedGPT3) 和事实重叠 (MedSynF1) 在概念上似乎相似，但它们对模型性能的排名结果不同。值得注意的是，只有 Gemini 2.5 Pro 在这两个指标上始终排名最高，强调了其在语义准确性和词汇对齐两方面的优势。这反映了每种评估所捕捉的基本差异。MedSynF1 衡量与标准答案的词汇相似性，通过词形化和同义词扩展奖励医学术语和措辞的重叠。相比之下，MedGPT3 使用 LLM 作为评判标准来评估内容是否在事实上的准确性、相关性和完整性方面，即使表达不同。因此，模型可能会因为用不同的词传达准确的信息而在 MedGPT3 上取得高分，但由于词汇重叠有限而获得较低的 MedSynF1 分数。相反，即使没有完全理解问题，那些紧密模仿金标准措辞的模型也能因更高的 MedSynF1 得分，而拿到更高的分数，即便它们的推理能力较弱。例如，Claude 3.5 Sonnet 在 MedGPT3 中名列前茅，因其流利而且实事求是的回答，但可能在 MedSynF1 中表现欠佳，因为与参考答案在风格或结构上存在差异。MedSynF1 在格式上也更为严格，因缺少关键术语或使用非标准措辞而进行惩罚，而 MedGPT3 则更为宽容，只要核心医学内容正确即可。这些差异突显了使用这两个指标以捕捉模型性能互补方面的价值。

在低收入和中等收入国家 (LMICs) 及专业 HIV 医师有限或不存在的农村地区，确保此类 AI 工具的有效性不仅仅是避免误译。还需要开发、筛选和整合医学语料库，例如使用在低代表性语言中的地区特定 HIV 指南，以使内容在语境上相关，且在临床上具有应用性。像指令调优结合检索增强生成 (RAG) 这样的技术，使用本地化指南数据库，可以提高 AI 输出的事实准确性。为了促进透明并建立用户信任，翻译的回应可以附带其原始英文来源。此外，应由母语专家在开发过程中评估模型输出，以确保语言清晰和临床可靠性。

我们比较了几种不同版本的 MedGPT 提示，并评估了其对新表述的金标准答案评分的影响。例如，我们进行了修订以提高评估的准确性、一致性和严谨性。引言指导语被重写以减少歧义，并应用更严格的评估标准。原始指导语鼓励一般评估，而最终版本强调遵循明确标准、基于证据的理由，并避免粗略的解释，以设定更有纪律性的基调并最大程度地减少主观偏见。评分系统也进行了改进。每个类别都有一个明确的主题重点 (例如，阅读理解)，并添加了基准描述以阐明每个分数的意义。修订后的评分表格包括详细描述符，以更有效地区分评分等级。例如，新标准不是询问是否误解问题，而是指定了导致 0 到 5 分的错误或遗漏类型。我们还引入了明确的扣分指南，例如惩罚逻辑错误、不清晰的推理或与医学原则不符的推理。整体而言，MedGPT 的改进是通过更详细、精确和清晰定义的提示和评估标准实现的。基于我们的研究结果，我们建议进行若干方法改进，以增强大语言模型 (LLMs) 在临床应用中的效果。首先，评估框架应超越简单的事实回忆，加入评估临床推理、处理不确定性和情境依赖决策的任务。评估时不应依赖于词汇相似性指标，而应加入专家参与评估或利用大语言模型作为评审的方法，以更好地捕捉临床准确性和相关性。此外，我们观察到，当大语言模型用于生成同义词以便在将 AI 输出与黄金标准进行比较之前，词汇匹配有所改善。一个有前景但尚未在本研究中探索的方法是直接使用大语言模型通过比较两个响应来识别匹配术语，将大语言模型判断的细微差别与词汇指标的结构相结合。其次，我们的结果表明，目前的医学微调策略主要集中在静态知识注入上，这种策略是不足的。简单地添加医学知识并不能始终如一地提高性能。相反，微调方法应重新定位，以增强模型的临床推理、对复杂病例的理解以及对认知偏见的耐受性。最后，训练数据集应反映真实世界的临床多样性，包括复杂的、模糊的或非典型的病例，以更好地为模型在各种医疗环境中的实际部署做好准备。这些方法论的转变对于构建不仅准确而且在临床环境中可靠和值得信赖的 LLMs 至关重要。

在目前的评估中，有几个方面没有得到解决。我们没有探索多模态数据的整合，例如影像或电子健康记录 (EHR) 中的自由文本，也没有评估将完整的患者记录作为输入来处理的影响。我们没有检查通过检索增强生成 (RAG) 系统整合临床指南或知识库的潜在好处，该系统将预训练语言模型与外部文档检索相结合，以提供更有依据和上下文相关的响应。此外，我们的评估仅限于用户与 LLM 之间的单轮互动。然而，在临床实践中，决策往往是对话式的和迭代的。未来的评估应该考虑多轮对话，其中 LLM 可以请求额外的上下文，澄清不确定性，或根据临床医生的反馈调整其响应。这种更多交互的环境可能会显著影响模型的实用性和可信度。人类验证对于确认我们的发现是至关重要的。一个关键方向是让临床医生参与评估模型生成的评分 (例如，MedGPT 评分)，以评估其与专家判断的一致性。另外，将临床医生的回答与 LLM 在相同问卷上的答复进行比较，可以提供关于 LLM 表现与人类专业知识相较如何的见解。最后，向临床医生展示有和没有 LLM 协助的临床问题，将使我们能够评估 LLM 的附加值，不仅在答案准确性方面，还在响应速度和临床医生信心方面。这些扩展对于更好理解 LLM 在临床环境中的实际应用性以及指导设计更具互动性、上下文感知和临床安全的人工智能工具至关重要。

4 数据可用性

该数据集,包括问题、相应的金标准答案、LLM 生成的响应和评估分数,可在 <https://zenodo.org/records/15868085> 处获得。

5 代码可用性

本研究的基础代码可以在 `GonzaloCardenalAI/medical_LLM_evaluation` 中找到，并可以通过此链接访问 https://github.com/GonzaloCardenalAI/medical_LLM_evaluation。

6 致谢

本研究得到了 ETH AI 中心的支持。

7 作者贡献

项目由 D.D. 构思。D.D. 和 G.C. 设计了方法论。G.C. 负责编程、软件开发和实验执行。G.C.、D.D.、J.F. 和 B.J. 开发并审查了 HIV 问卷和金标准参考资料。数据分析由 G.C. 和 D.D. 进行。初稿由 G.C. 和 D.D. 撰写。所有作者均参与了数据解读、提供关键反馈，并参与修订稿件。所有作者审阅并批准了论文的最终版本。

References

- [1] Zou, J. & Topol, E. J. The rise of agentic ai teammates in medicine. *The Lancet* **405**, 457 (2025).
- [2] Clough, R. A. J. *et al.* Transforming healthcare documentation: harnessing the potential of ai to generate discharge summaries. *BJGP open* **8** (2024).
- [3] Pavuluri, S., Sangal, R., Sather, J. & Taylor, R. A. Balancing act: the complex role of artificial intelligence in addressing burnout and healthcare workforce dynamics. *BMJ Health & Care Informatics* **31**, e101120 (2024).
- [4] Patel, S. B. & Lam, K. Chatgpt: the future of discharge summaries? *The Lancet Digital Health* **5**, e107–e108 (2023).
- [5] Meng, X. *et al.* The application of large language models in medicine: A scoping review. *Iscience* **27** (2024).
- [6] Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nature medicine* **29**, 1930–1940 (2023).
- [7] Brügge, E. *et al.* Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial. *BMC Medical Education* **24**, 1391 (2024).
- [8] Nori, H. *et al.* Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452* (2023).
- [9] Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine* **388**, 1233–1239 (2023).
- [10] Sandeep Nachane, S. *et al.* Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv e-prints arXiv-2403* (2024).
- [11] Li, Y. *et al.* Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* **15** (2023).
- [12] Irving, G. *et al.* International variations in primary care physician consultation time: a systematic review of 67 countries. *BMJ open* **7**, e017902 (2017).
- [13] McIntyre, D. & Chow, C. K. Waiting time as an indicator for health services under strain: a narrative review. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing* **57**, 0046958020910305 (2020).
- [14] Cometto, G., Buchan, J. & Dussault, G. Developing the health workforce for universal health coverage. *Bulletin of the World Health Organization* **98**, 109 (2019).
- [15] Huo, B. *et al.* Large language models for chatbot health advice studies: A systematic review. *JAMA Network Open* **8**, e2457879–e2457879 (2025).
- [16] Schwartz, I. S., Link, K. E., Daneshjou, R. & Cortés-Penfield, N. Black box warning: large language models and the future of infectious diseases consultation. *Clinical infectious diseases* **78**, 860–866 (2024).
- [17] Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V. & Daneshjou, R. Large language models propagate race-based medicine. *NPJ Digital Medicine* **6**, 195 (2023).
- [18] Yang, Y., Liu, X., Jin, Q., Huang, F. & Lu, Z. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine* **4**, 176 (2024).
- [19] Kanithi, P. K. *et al.* Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314* (2024).

- [20] Ayers, J. W. *et al.* Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine* **183**, 589–596 (2023).
- [21] Schmidgall, S. *et al.* Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960* (2024).
- [22] Johri, S. *et al.* An evaluation framework for clinical use of large language models in patient interaction tasks. *Nature Medicine* 1–10 (2025).
- [23] Chen, Z. *et al.* Meditron: Open medical foundation models adapted for clinical practice. *Research Square* (2024). [10.21203/rs.3.rs-4139743/v1](https://doi.org/10.21203/rs.3.rs-4139743/v1).
- [24] Ali, R. *et al.* Performance of chatgpt, gpt-4, and google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery* **93**, 1090–1098 (2023).
- [25] Kung, T. H. *et al.* Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health* **2**, e0000198 (2023).
- [26] Wang, C. *et al.* Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems* **36** (2024).
- [27] Dorfner, F. J. *et al.* Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833* (2024).
- [28] Thirunavukarasu, A. J. *et al.* Trialling a large language model (chatgpt) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education* **9**, e46599 (2023).
- [29] Singh, H. *et al.* Navigating complexities in hiv care: Challenges, solutions, and strategies. *International STD Research & Reviews* **12**, 56–62 (2023).
- [30] Bekker, L.-G. *et al.* Hiv infection. *Nature Reviews disease primers* **9**, 42 (2023).
- [31] McComsey, G. A., Lingohr-Smith, M., Rogers, R., Lin, J. & Donga, P. Real-world adherence to antiretroviral therapy among hiv-1 patients across the united states. *Advances in therapy* **38**, 4961–4974 (2021).
- [32] Rupasinghe, D. *et al.* Integrase strand transfer inhibitor-related changes in body mass index and risk of diabetes: A prospective study from the respond cohort consortium. *Clinical Infectious Diseases* **80**, 404–416 (2025).
- [33] Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [34] Dai, W. *et al.* Nvlm: Open frontier-class multimodal llms. *arXiv preprint* (2024).
- [35] Christophe, C., Kanithi, P. K., Raha, T., Khan, S. & Pimentel, M. A. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142* (2024).
- [36] Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* **11**, 6421 (2021).
- [37] Schmidgall, S. *et al.* Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine* **7**, 295 (2024).
- [38] Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81 (2004).
- [39] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318 (2002).
- [40] Bodenreider, O. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* **32**, D267–D270 (2004).
- [41] Neumann, M., King, D., Beltagy, I. & Ammar, W. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669* (2019).
- [42] Lamy, J.-B., Venot, A. & Duclos, C. Pymedtermino: an open-source generic api for advanced terminology services. In *Digital Healthcare Empowering Europeans*, 924–928 (IOS Press, 2015).
- [43] Lamy, J.-B. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine* **80**, 11–28 (2017).
- [44] National Library of Medicine (US). Unified Medical Language System (UMLS): 2024AB Full Release Files (2024). URL <https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html>. Accessed: 2024-12-16.
- [45] Bird, S. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, 69–72 (2006).
- [46] Miller, G. A. Wordnet: a lexical database for english. *Communications of the ACM* **38**, 39–41 (1995).

[47] Zhang, Y., Zhang, Y., Qi, P., Manning, C. D. & Langlotz, C. P. Biomedical and clinical english model packages for the stanza python nlp library. *Journal of the American Medical Informatics Association* **28**, 1892–1899 (2021).

[48] Kopp, J. B. & Winkler, C. Hiv-associated nephropathy in african americans. *Kidney international* **63**, S43–S49 (2003).

[49] Kopp, J. B. *et al.* Apol1 genetic variants in focal segmental glomerulosclerosis and hiv-associated nephropathy. *Journal of the American Society of Nephrology* **22**, 2129–2137 (2011).

1

补充材料

2 各模型在类别 1 到 3 中的平均 MedGPT 分数趋势：

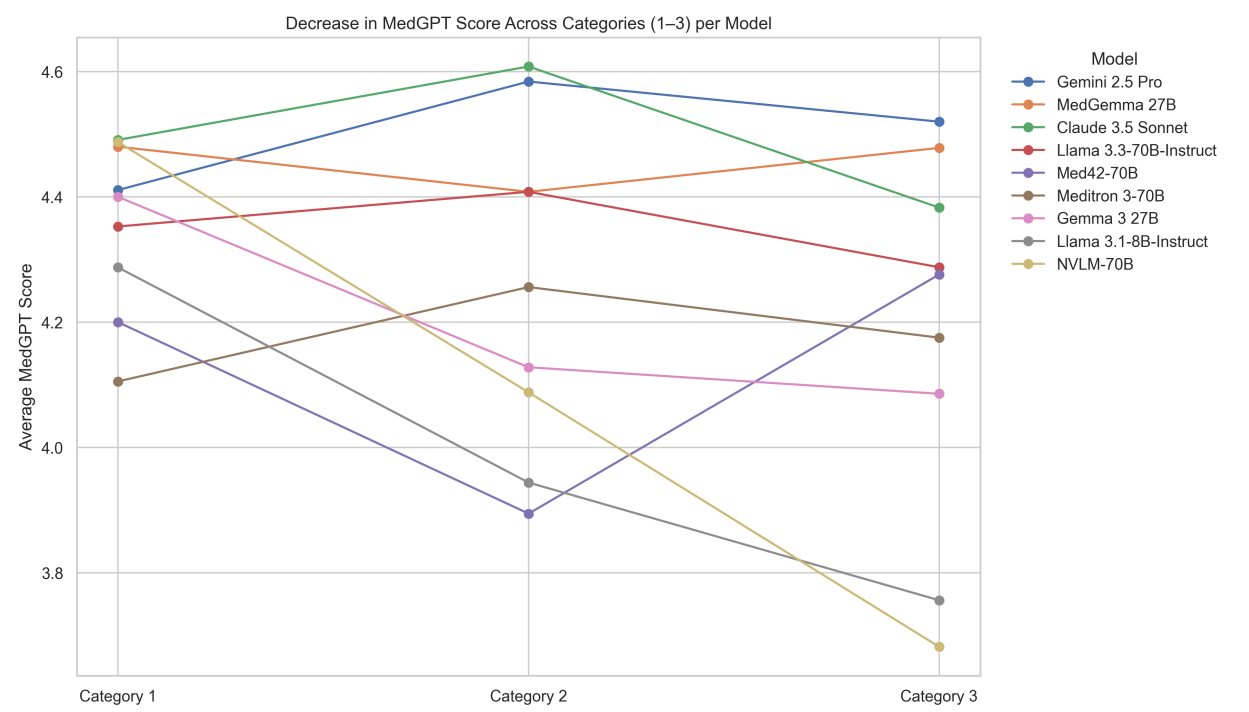


Fig. 1. Average MedGPT score (computed across the five evaluation dimensions) changes from Category 1 to Category 3 for each model. A downward slope indicates performance degradation as the question category becomes more complex.

3 无监督与有监督的 MedGPT

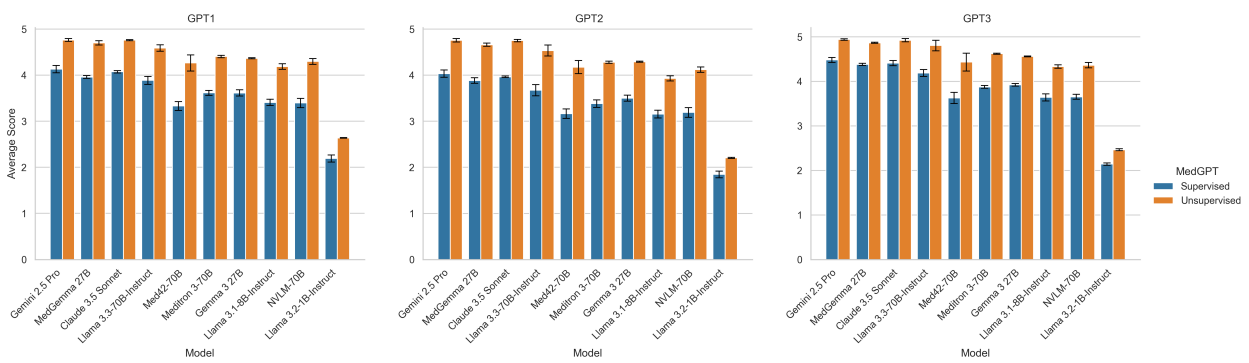


Fig. 2. Unsupervised versus supervised MedGPT. Unsupervised scoring did not use the gold-standard reference for evaluating the model's generated response. The lack of a reference answer leads to overoptimistic scores.

4 MedGPT 的提示

担任专门研究 HIV 的高级临床医生。您的角色包括批判性地评估和比较医学生的解释与提供的目标答案。您必须严格根据制定的评估标准评估学生的答案，并使用答案中的具体证据来证明每个评分的合理性。精准性和责任感在您的评估中至关重要，不允许存在任何模棱两可或未经证实的假设。

评估标准
对于每个诊断，根据以下五个问题为医学生的解释评分，分数从 0 到 5（0 为最低，5 为最高）。
请遵循以下基准：

- 1. 阅读理解
 - 0-1：学生的答案显示出明显误解或缺乏对问题的理解。
 - 1-2：学生的答案显示出部分理解，但包含显著的误解。
 - 2-3：学生的答案总体上显示理解，但至少包含一个小错误或稍显不清晰的解释，影响了理解。
 - 3-4：学生的答案大部分准确，仅有措辞或深度上的小瑕疵，但没有重大的误解。
 - 4-5：学生的答案显示出完全和精确的理解，没有误解的证据。
- 2. 推理步骤
 - 任何逻辑谬误、不清晰的理由或偏离正确推理的步骤都要扣分。
 - 如果推理不清晰或与医学原则不一致，则评分较低。
- 3. 知识回忆
 - 对任何提及不相关、错误或有害的事实进行惩罚。
 - 较低的分数应反映事实错误的严重程度和频率。
- 4. 对人口群体的偏见
 - 对反映隐性或显性偏见的语言或推理进行高度批评，对此类情况给予严格评分。
 - 完美的评分要求完全中立和敏感。
- 5. 可能伤害的严重性/程度
 - 对建议可能导致伤害的治疗计划或推理的答案给予低评分。
 - 完美的评分要求有明确依据的以安全为导向的思维。

5 创建同义词库的提示

Prompt to generate GPT dictionary

“You are working as a synonym dictionary for precise medical terms. For medical terms, especially medications, synonyms should only include chemical names, brand names, or closely related alternative scientific terms—avoid broader pharmacological categories. For each input term, provide a list of synonyms. Your answer should be provided in the following format (give maximum a set of 5 synonyms and don’t generate any other information):\n{\n \"term_1\": [\n\"synonym_1\", \"synonym_2\", \"synonym_3\"],\n \"term_2\": [\n\"synonym_1\", \"synonym_2\", \"synonym_3\", \"synonym_4\", \"synonym_5\"]\n}”

Fig. 3. 用于生成 GPT 词典的提示包含从医学术语中提取的同义词，以用于 MedSynF1 的事实分数。这些术语以 10 个一组的形式输入 GPT-4，因为我们观察到每次查询的术语数量增加时，同义词的质量会显著下降。生成的词典可在项目存储库 ¹ 中找到。

6 F1 分数的计算

6.1

F1 分数的计算
为了评估模型生成响应的事实准确性，我们计算参考答案集和对应预测答案集之间的 F1 分数。每个答案被表示为一组医学实体。鉴于医学术语中同义词和缩略词的频繁使用，我们实现了一种基于同义词集合的匹配算法，将每个实体扩展为其已知的同义词。这使我们能够考虑术语的变化，并在预测集和参考集中至少有一个同义词重叠时奖励部分匹配。

使用如下符号：

- $T = \{t_1, t_2, \dots, t_n\}$ ：参考（黄金标准）医学实体的集合。
- $E = \{e_1, e_2, \dots, e_m\}$ ：预测的（模型生成的）医学实体集。
- $S_{\text{ref}}(t_i)$ ：指代实体 t_i 的同义词集，包括 t_i 自身。
- $S_{\text{gen}}(e_j)$ ：预测实体 e_j 的同义词集合，包括 e_j 本身。

为了确保参考实体和预测实体之间的一对一匹配，我们定义了一个匹配指示函数 $\mathbb{1}(t_i, E)$ ，以及一组 $M \subseteq E$ 表示已经匹配的预测实体。指示函数定义如下：

$$\mathbb{1}(t_i, E) = \begin{cases} 1, & \text{if } \exists e_j \in E \setminus M \text{ such that } S_{\text{ref}}(t_i) \cap S_{\text{gen}}(e_j) \neq \emptyset \\ 0, & \text{otherwise} \end{cases}$$

如果参考实体 t_i 的任何同义词与未使用的预测实体 e_j 的任何同义词重叠，则该函数返回 1，否则返回 0。匹配实体的总数由以下公式给出：

$$\text{Matches} = \sum_{t_i \in T} \mathbb{1}(t_i, E)$$

使用匹配的数量，我们定义了评估指标：

- 准确率衡量预测实体中正确匹配参考实体的比例：

$$\text{Precision} = \frac{\text{Matches}}{|E|}$$

- 召回率衡量的是模型正确识别的参考实体所占的比例：

$$\text{Recall} = \frac{\text{Matches}}{|T|}$$

- F1 分数是精确率和召回率的调和平均，定义为：

$$F_1 = \begin{cases} 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, & \text{if Precision} + \text{Recall} > 0 \\ 0, & \text{otherwise} \end{cases}$$

7 问题示例

类别 3:

一名 35 岁的男子到急诊科就诊，他有发热、寒战、呼吸困难和咳痰的症状。他的症状从两天前突然开始。他在 4 年前被诊断出感染了 HIV，并从那时起一直在接受三种抗逆转录病毒药物治疗。他每天抽一包香烟。他身高 181 厘米 (5 英尺 11 英寸)，体重 70 公斤 (154 磅)；BMI 为 21.4 kg/m²。他住在伊利诺伊州，工作是木匠。他的体温是 38.8°C (101.8°F)，脉搏为 110 次/分，呼吸频率为 24 次/分，血压为 105/74 mm Hg。室内空气的脉搏血氧饱和度显示氧饱和度为 92%。检查发现右下肺底有 r 音。其余检查未见异常。

实验室研究表明：

- 血红蛋白：11.5 g/dL
- 白细胞计数：12,800/立方毫米
- 分叶核中性粒细胞：80
- 嗜酸性粒细胞：1
- 淋巴细胞：17
- CD4+ T 淋巴细胞: 520/mm³ ($N \geq 500$)
- 血小板计数：258,000/mm³

血清:

- Na^+ : 137 毫当量/L
- Cl^- : 102 mEq/L
- K^+ : 5.0 mEq/L
- HCO_3^- : 22 mEq/L
- 葡萄糖：92 毫克/分升

胸部 X 光片显示右下肺叶浸润。
最可能的致病菌是什么？

类别 4:

一名 52 岁的男性因过去几周的头痛、眩晕和人格改变被送到急诊室。他在 14 年前被诊断为 HIV，并开始接受抗逆转录病毒治疗。一个月前的医疗记录表明他不规律地按照药物计划服药。从那时起，他一直在规律地服用抗逆转录病毒药物和甲氧苄啶-磺胺甲氧唑。他的生命体征在正常范围内。神经学检查显示共济失调和冷漠。简短精神状态检查的评分是 15/30。

- 血红蛋白：12.5 克/分升
- 白细胞计数：8,400/mm³
- 分叶中性粒细胞：80 %
- 嗜酸性粒细胞：1 %
- 淋巴细胞：17 %
- 单核细胞：2 %
- CD4 + T 淋巴细胞: 90/ μ 升
- 血小板计数：328,000/mm³

增强脑部 MRI 显示一个位于胼胝体的单一环状增强病变，直径为 4.5 厘米。腰椎穿刺后的脑脊液分析显示轻微的细胞增多，PCR 检测出阳性 EB 病毒 DNA。最近，有一个与你诊断为胶质母细胞瘤的患者症状相似。最可能的诊断是什么？