# RARL: 在数据和硬件限制下,通过强化学习和 LoRA 改善医学 VLM 推理和泛化

**Tan-Hanh Pham**<sup>1,2,†</sup>, **Chris Ngo**<sup>2</sup>

<sup>1</sup>Harvard Medical School, USA, <sup>2</sup>Knovel Engineering Lab, Singapore

<sup>†</sup>Corresponding author

## Abstract

视觉语言模型(VLMs)在医疗应用中的日益整合为诊断推理提供了有前 途的支持。然而,目前的医疗 VLMs 往往在泛化性、透明性和计算效率方 面存在限制,这些障碍阻碍了它们在资源受限的实际环境中的部署。为了 解决这些挑战,我们提出了一种推理感知强化学习框架 RARL,该框架增 强了医疗 VLM 的推理能力,同时保持了在低资源环境中的效率和适应性。 我们的方法通过低秩适配和定制奖励函数微调了一个轻量级的基础模型 Qwen2-VL-2B-Instruct,这些奖励函数共同考虑诊断准确性和推理质量。训 练是在一台 NVIDIA A100-PCIE-40GB GPU 上进行的,展示了在受限环境 中部署此类模型的可行性。我们使用 LLM-as-judge 框架评估模型,该框架 对正确性和解释质量进行评分。实验结果表明, RARL 显著提高了医学图 像分析和临床推理中的 VLM 性能, 在侧重于推理的任务上比监督微调提高 了约7.78%,同时需要更少的计算资源。此外,我们展示了该方法在未见 数据集上的泛化能力,比监督微调提高了约27%的性能,并比传统的RL 微调高出约4%。我们的实验还表明,训练期间的多样性提示和推理期间 的推理提示对于提升 VLM 性能至关重要。我们的研究结果强调了推理引导 学习和推理提示在引导医疗 VLMs 走向更透明、准确且资源高效的临床决 策过程中具有的潜力。代码 和 数据 是公开可用的。

# 1 引言

大型语言模型(LLMs)的快速进展显著地重塑了人工智能(AI),使机器能够以惊人的流利 程度执行复杂的语言理解和生成任务[1-5]。这一进步为多模态模型的发展奠定了基础,尤 其是结合视觉感知与语言推理的视觉-语言模型。这些创新引发了在视觉和文本数据自然交 叉的领域中日益增长的兴趣,尤其是在医学领域。在临床环境中,结合诊断图像与文本患者 数据进行解读是有效决策的核心。因此,医学视觉语言模型作为支持放射科医生和临床医 生的有前景工具出现,能够执行自动异常检测、报告生成和基于图像的问答等任务[6-10]。 例如,像 Llava-Med [11]和 Med-PaLM M [8]这样的模型已展示出分析胸部 X 光片识别病变 或从超声图像生成结构化报告的能力。这些进展对增强临床工作流程具有显著的潜力,尤 其是在快速和准确诊断至关重要的高风险环境中。

尽管在这些方面取得了进展, 医疗大模型(VLMs)的发展仍受到若干实际挑战的限制。首先, 高性能的模型, 如 LLaVA-Med [11] 或 Med-PaLM M [8], 依赖于大规模数据集(例如, MIMIC-CXR [12] 或 PMC-15M [13])和大量的计算资源,通常需要集群的高端 GPU 进行训练。这些要求对较小的医疗机构、资源匮乏的临床环境或全球健康计划构成了巨大阻碍 [14]。其次,许多医疗大模型针对特定任务或数据集进行了优化,导致在未见过或多样化的临床场景中表现不佳。例如, 一个在单一医院的胸部 X 光片上训练的模型可能难以分析来自不同成像协议或患者人群的图像 [15,16]。然而,尽管这些系统具有潜力,仍然存在若干挑战。值得注意的是,许多现有模型在跨多样化医疗数据集的泛化性上表现挣扎,需要计算密集

Preprint. Under review.

型的训练机制,而最为关键的是,它们缺乏透明的推理能力,这对临床信任和责任追踪至关 重要[17]。

为了解决这个问题,近来的研究探索了以推理为中心的技术,例如链式思维(CoT)提示, 鼓励模型在给出答案之前表达中间步骤[18]。例如,CoT已被应用于视觉语言模型(VLMs) 中,以通过生成结构化的诊断解释提升在医学问答任务中的表现,如在得出诊断结论之前 识别放射学征象[19]。对于数据集,医学推理数据集也被开发来增强医学 VLMs的推理能 力,例如[20]。在训练技术方面,强化学习(RL)在增强医学 VLMs的推理能力上相对于 自监督学习展示了其潜力,这得益于 Deepseek 的成功[21]。

最近,基于强化学习(RL)的方法在优化诊断准确性和推理深度方面受到关注[22]。像 Med-R1[16]和 MedVLM-R1[15]这样的研究通过奖励驱动的优化改进模型输出,在检测胸 片异常等任务中提高了性能。一个关键的创新是群体相对策略优化(GRPO),这是一种先 进的 RL 框架,能够平衡包括诊断正确性、解释清晰度和响应简洁性在内的多个目标[23]。

尽管做出了这些努力, 医学 VLM 仍然局限于多项选择或封闭式问题; 除此之外, 这些模型 在泛化到其他类型的问题或分布外的数据时存在困难。此外, 尽管 MedVLM-R1 [15] 使用强 化学习(RL)进行训练, 但它在不同问题类型之间的泛化能力不佳。此外, 尽管最近的进 展改善了医学 VLM 的可解释性, 但其在资源受限环境中的应用仍未得到充分探索。在评估 方面,目前评估推理模型的方法主要集中在最终答案的准确性上,常常掩盖了推理过程, 使 得模型"准备不足", 无法提供临床验证所需的详细解释 [17]。

为了解决这些限制,我们提出了一种资源高效的训练流程,在有限的数据和硬件条件下增强 推理和诊断能力。我们的方法结合了一种新颖的推理感知强化学习(RARL)策略与低秩适应 (LoRA)微调。利用 Qwen2-VL-2B 这款轻量级 VLM,我们在单个 NVIDIA A100-PCIE-40GB GPU 上训练我们的模型。在评估时,我们采用一个 LLM 作为裁判的框架,全面评估答案的 正确性和推理质量。

我们的贡献总结如下:

- 我们引入了 RARL,这是一种强化学习框架,旨在有效提高小型 VLMs 在资源限制 下的推理质量、可解释性和泛化能力,从而解决医疗 AI 部署中的一个关键缺口。
- 在单个 GPU 上训练证明了其在真实临床环境下的可行性。
- 在未见过的数据集(例如, VQA-RAD)上获得 27 % 的增益, 突显了框架的适应性。

# 2 方法论

#### 2.1 数据集

为了弥合通用视觉语言理解与特定领域医学推理之间的鸿沟,我们假设明确的推理监督是 必不可少的。为支持这一点,我们从 Silvar-Med [20] 中精选了一小部分优质推理数据集,该 数据集包含由三名医生标注的 866 个样本,涵盖 MRI (22.4%,194 个样本)、CT 扫描 (16.5%,143 个样本)和胸部 X 光片 (61.1%,529 个样本)。该数据集分为训练集和测试集:716 个用于训练 (162 个 MRI,122 个 CT,432 个 X 光),150 个用于测试 (32 个 MRI,21 个 CT, 97 个 X 光)。与之前主要专注于预测简短最终答案或多项选择题 [24] 的工作不同,我们的研 究明确地实验了推理可解释性。在该数据集中,每个问题都配有一个详细的基于推理的答 案和原始答案,模型不仅要预测正确的标签,还必须提供支持决策的临床一致的解释。数据 样本如表 1 所示。除了我们精选的数据集,我们还使用 VQA-RAD [25]、SLAKE (英语) [26] 、VQA-Med 2019 [27] 评估我们方法的泛化能力。

我们使用 GRPO 微调模型,这是一种强化学习算法,旨在通过为每个输入采样多个输出并 基于组归一化相对优势 [28] 进行优化来增强推理能力。GRPO 通过直接从采样的奖励中估 算基准来消除对单独价值网络的需求,从而降低资源受限环境下的计算开销。

给定一个输入问题 q , 我们从旧策略  $\pi_{\theta_{old}}$  中抽样出一组 G 候选输出  $\{v_1, v_2, \ldots, v_G\}$  。然后 通过最大化以下目标来优化策略  $\pi_{\theta}$ :

Sample 1: Abnormality detection	Sample 2: Abnormality detection
Question : Does this image look abnormal?	Question : Is the lung healthy?

Answer : <think> Yes, the image appears abnormal due to the presence of irregularities in the lung fields and potential signs of pathology. The overall lung structure may show unusual densities or patterns that deviate from normal anatomical features. These findings could indicate conditions such as infection, inflammation, or other pulmonary issues that warrant further investigation





Explainable Answer : <think> The lung appears to show signs of pathology, indicated by the presence of abnormal structures and potential lesions. These findings suggest that the lung is not functioning optimally and may be affected by a disease process. Further evaluation and clinical correlation are necessary to determine the specific condition affecting the lung. Overall, the imaging suggests that the lung is not healthy

Short answer : No.



$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim p(Q), \{v_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | q)} \left[ \frac{1}{G} \sum_{i=1}^G \min\left(r_i(\theta) A_i, \operatorname{clip}(r_i(\theta), 1 - \epsilon, 1 + \epsilon) A_i\right) - \beta \operatorname{D}_{\text{KL}}\left(\pi_{\theta}(\cdot | q) \| \pi_{\text{ref}}(\cdot | q)\right) \right],$$

$$(1)$$

,其中优势 A<sub>i</sub> 是从一组奖励 {r<sub>1</sub>, r<sub>2</sub>,..., r<sub>G</sub>} 计算得出的:

$$A_{i} = \frac{r_{i} - \text{mean}(\{r_{1}, r_{2}, \dots, r_{G}\})}{\text{std}(\{r_{1}, r_{2}, \dots, r_{G}\})},$$
(2)

而 KL 散度项是:

$$D_{KL}(\pi_{\theta} \parallel \pi_{ref}) = \frac{\pi_{ref}(v_i \mid q)}{\pi_{\theta}(v_i \mid q)} - \log \frac{\pi_{ref}(v_i \mid q)}{\pi_{\theta}(v_i \mid q)} - 1.$$
(3)

在此,  $\epsilon$  控制剪裁范围以保持稳定性,  $\beta \in$  KL 惩罚系数,用于防止与参考策略 π<sub>ref</sub>(预训练 模型)过度偏离。 奖励  $r_i$  是通过一个组合规则函数计算的,该函数评估输出格式(如 <think> 和 <answer> 标签的正确使用)、初始推理的正确性以及最终预测。

我们设计了奖励函数,该函数由多个组件组成,以促进临床正确性和可解释的推理。

格式奖励:此项通过检查 <think>...</think> 和 <answer>...</answer> 标签的存在来强化输出 结构。如果输出完全符合格式,则给予全额奖励(1.0)。如果只有部分标签存在,则给予小 额奖励(0.1)。没有任何标签的输出则不获得奖励。该部分确保模型始终产生可解释且结构 化的响应。

长度奖励:这一术语通过奖励较长的输出来激励提供足够详细的解释。奖励与输出长度成正比,以标记数来衡量,线性缩放至最大上限1.0。具体而言,分配min(0.001×token count,1.0)的奖励。这可以防止缺乏足够推理的极短或琐碎的完成。

准确性奖励:这一项衡量模型最终答案的正确性,该答案被包含在 <answer>...</answer>标记内。对于多项选择题或二选一(是/否)问题,给予二进制奖励(对为1.0,错则为0.0)。对于开放性问题,使用 BERTScore F1 [29] 来计算预测答案和参考答案之间的语义相似性,从而产生一个介于 0.0 到 1.0 之间的连续奖励。

推理奖励:除了上述的奖励之外,我们引入了一种推理感知奖励,专门激励在 <think>...

 <think>...
 标签内生成有意义的结构化推理。这种奖励的动机源于我们的观察:基础 模型在通用领域数据上进行预训练,具有有限的医学知识,因此需要额外的指导才能形成 连贯的临床推理路径。这一观察类似于现实情况:当识别大脑的 MRI 图像时,你知道某些 特征暗示肿瘤而不是良性囊肿,但不懂得如何逻辑地解释原因。如果没有明确的推理指导, 模型(甚至是人类学习者)可能会记住视觉模式("亮点=肿瘤"),而无法理解如肿块效应、 周围水肿或对比增强模式等关键的潜在特征。因此,为了在适应医学领域的过程中培养临 床上有意义的推理,而不是肤浅的模式识别,激励结构化思维过程是至关重要的。

#### 2.2 低秩适应

为了进一步减少硬件资源的训练参数数量,我们利用低秩适应微调。LoRA 被应用于秩 r = 8 、 α = 16 和目标模块 ["q\_proj", "k\_proj", "v\_proj", "o\_proj"] 来减少可训练参数, 从而实现单 GPU 训练。视觉编码器被冻结,以将优化重点放在语言组件上,从而减少计算 开销。我们训练模型 5 个世代,每 100 步保存检查点,以捕捉峰值性能并减轻不稳定性。此 外,我们使用 Flash Attention 来加速更快的训练和推理 [30]。

在将我们提出的训练流程应用到医学领域之前,我们首先使用 Cambrian Vision-Centric Benchmark (CV-Bench) 来验证其在通用领域任务上的有效性。这一步的动机是因为基础模型 在通用领域数据上进行了广泛的预训练,因此评价其性能变化是合适且公正的。CV-Bench 评估多样的视觉推理技能,如物体计数、空间关系推断和深度排序,使其成为一个理想的基准,用于单独分析我们的训练流程的效果,而无需考虑领域适应的复杂性。我们的流程结 合了用于参数高效微调的 LoRA 和基于 GRPO 的强化学习。需要注意的是,在这次实验中我们没有使用推理意识奖励,因为数据集中缺乏推理注释。为了评估其在数据稀缺条件下的鲁棒性,我们使用 500、1000 和 5000 个训练样本对模型进行微调,并比较三种训练设置:(1)标准监督微调 (SFT),(2) 不使用 LoRA 的基于 GRPO 的强化学习,以及(3) 我们提出的方法:使用 LoRA 的基于 GRPO 的强化学习。

Method	Training Samples	Count Acc. (%)	Relationship Acc. (%)	Depth Acc. (%)	Distance Acc. (%)	Total Acc. (%)
Experiment from literatures						
RL only (w/o LoRA) [32]	15,000	69.54 %	61.84 %	66.50 %	65.50 %	66.03 %
Our experiments with Qwen2-VL-2B-Instruct						
SFT	500	59.26 %	52.61 %	34.34 %	38.33 %	47.19 %
SFT	1,000	59.90 %	51.54 %	46.67 %	46.00 %	51.67 %
SFT	5,000	60.27 %	65.54 %	45.67 %	42.83 %	54.28 %
RL w/ LoRA	500	58.74 %	54.19 %	38.00 %	41.23 %	48.04 %
RL w/ LoRA	1,000	60.30 %	57.23 %	48.33 %	47.47 %	53.33 %
RL w/ LoRA	5,000	63.89 %	68.61 %	52.84 %	50.83 %	59.04 %
RL only (w/o LoRA)	5,000	64.43 %	65.65 %	50.35 %	55.71 %	59.29 %

Table 2: 在不同样本规模下比较 CV-Bench 上的不同训练策略。"RL w/ LoRA"在多种数据规模下进行评估以展示可扩展性。

结果如表2所示,其中我们在训练过程中使用相同的提示进行评估(推理提示)。实验表明, 在较小的训练数据集(500和1,000个样本)下,我们的方法(RL with LoRA)在 CV-Bench 上优于 SFT 方法,表明在限制数据场景中的性能更佳。此外,当使用较大数据集(5,000个 样本)进行训练时,使用 RL 和 LoRA 的模型性能依然与未使用 LoRA 训练的模型具有竞争 力。这表明我们的方法不仅对较小的数据集有效,而且在数据可能受限的现实场景中也很 有前景,展示了其可扩展性和鲁棒性。

## 2.3 从一般领域到医学领域的适应

在研究了我们训练流程在 CV-Bench 上的效果之后,我们将训练流程扩展至医学领域。具体来说,我们在精心整理的以推理为重点的医学训练数据集上对模型进行微调,并在我们的 推理测试集和几个广泛使用的医学 VQA 基准测试上测试其性能: VQA-RAD [25]、SLAKE (英语) [26] 和 VQA-Med 2019 [27]。

与我们的测试集不同,其他基准测试主要由开放式或封闭式格式的非推理任务组成。由于 模型生成的自由格式输出与标准答案之间可能存在不匹配,直接的字符串匹配是不足够的, 因此我们采用 LLM-as-Judge 策略,包括 GPT-40 mini 和 Gemini 1.5 Flash,来评估模型生成 答案相对于标准答案的推理和准确性。对于推理,我们仅在我们自己的数据集上评估模型 预测与标准答案的对比,因为其他数据集不包含解释。

d

Dataset	GPT-40 mini		Gemini 1.5 Flash		Human Evaluation		
	Reasoning (%)	Final (%)	Reasoning (%)	Final (%)	Reasoning (%)	Final (%)	
Our experiment using SFT							
Ours	67.57 %	64.86 %	65.54 %	62.21 %	63.52 %	60.81 %	
VQA-RAD		26.16 %		22.37 %			
SLAKE		43.14 %		42.89 %			
VQA-Med 2019		13.80 %		11.43 %			
Path-VQA		9.09 %		8.88 %			
Our experiments with the propsed pipeline: RL + LoRA							
Ours	68.91 %	61.49 %	66.89 %	60.81 %	64.86 %	62.84 %	
VQA-RAD		43.90 %		42.13 %			
SLAKE		51.64 %		52.02 %			
VQA-Med 2019		49.80 %		49.20 %			
Path-VQA		24.26 %		23.05 %			

Table 3: 在不同微调方法下对不同医学 VQA 数据集上的模型进行评估。

表格 3 展示了模型在多个医学 VQA 数据集上的结果。我们比较了我们的管道与 SFT 微调 方法之间的结果。实验表明,基于我们测试集上的人工评估,模型在推理准确度上达到了 64.86%,在最终答案准确度上达到了 62.84%,展示了模型遵循逻辑推理过程的能力。此外, GPT-40 mini 和 Gemini 1.5 Flash 都表现出一致的评分趋势,验证了使用 LLM-as-Judge 框架 进行语义评估的有效性。值得注意的是,我们观察到推理分数与最终答案分数之间存在差 距。这一差异的出现是因为模型经常表现出有效的推理——包括正确的中间步骤——但可 能仍然产生与事实不完全匹配的答案。如表格 4 所示,模型的预测能够准确识别可能影响右 肺的潜在情况,包括肺炎,符合事实描述,但它也在没有充分理由的情况下提出了其他可能 性,如肺癌,稍微降低了其相关性。这强调了不仅需要评估最终答案,还要评估其背后的推 理过程的重要性。

为了泛化,表格3表明我们的流水线在各个公共基准上稳定地优于 SFT,在包括 VQA-RAD、 SLAKE 和 VQA-Med 2019 在内的准确性上提升了 5-35%。最低性能出现在 Path-VQA 数据 集上,该数据集是一个分布之外的大规模数据集,具有高度多样化的视觉和问题内容。表格 5提供了一个定性的例子,说明模型在 SLAKE 基准上为正常肝脏病例生成了准确且解释良 好的预测。

尽管我们的方法在医学 VQA 数据集上表现良好,但我们观察到了两个关键限制:

- 幻觉或虚弱的推理:模型有时会生成通用或依据不足的解释,这表明需要在问题与 推理之间更好地对齐。
- 简短回答的挑战:对于二元或事实性问题,模型可能会过度或不足地产生内容,难以在简洁和提供充足背景之间取得平衡。

## Table 4: 模型预测与推理。

Comparison between final prediction and the model reasoning.

Question : What disease is/are shown on the right of lung?

#### Short answer: 肺炎

Grounth truth: The image shows 肺炎 in the right lung, characterized by areas of increased opacity due to fluid accumulation or inflammation. This condition can result from various infectious agents, leading to symptoms such as cough, fever, and difficulty breathing. The presence of such opacities indicates that the lung tissue is affected, which is typical in pneumonia cases.



Model reasoning : The image is a chest X-ray, and the right lung appears to have a darker, more opaque area compared to the left lung. This is indicative of a lung opacity, which can be caused by various conditions such as 肺炎, 肺癌或其他肺部疾病。 It is important to consult with a healthcare professional for a proper diagnosis and treatment plan.

Model final prediction : pneumonia, lung cancer and other lung diseases.

## Table 5: 使用我们的流程对未见数据集进行模型预测。

Image	Explanation
	Question : Does the 肝脏 look normal? Ground truth : Yes. Model prediction : 肝脏表现出正常的大小和形状, 周围结构中没有 可见的病变或异常。The internal texture seems consistent, indicating healthy liver tissue. Overall, there are no signs of cirrhosis, fatty liver, or other pathological changes. Therefore, the liver can be considered normal based on this imaging.

Observation 2 : (1) The proposed method outperforms the SFT fine-tuning approach in both human and LLM-as-judge evaluations on the reasoning dataset. (2) The model demonstrates better generalization to unseen datasets when trained with our pipeline under limited data conditions. (3) We show that evaluating only the final answer is insufficient, as the model may perform meaningful reasoning but still fail to produce the optimal answer.

# 2.4 思维初始化和提示多样性



Figure 1: RARL 概述,它激励临床可信的中间推理和准确的最终答案。

如在第??和第2.3节所示,我们的小规模医学 VLM 管道在有限的数据和硬件限制下表现 优异。然而,我们注意到,这些模型在复杂的医学场景中仍倾向于产生幻觉或浅显的答案, 这表明仅针对最终答案准确性进行优化对于临床可靠性是不够的。为应对这一挑战,我们 引入了两个互补策略:推理感知强化学习(RARL)和提示多样性,旨在提高模型的推理一 致性、临床稳健性以及在资源有限的环境下适应多种医学查询的能力。

在 RARL 框架中,对于带有推理标注的训练样本,引导模型在生成最终答案之前,生成包含 在 <think> ... </think> 标签中的结构化中间思维。第?? 节提出的推理奖励用于评估这些 中间步骤的临床连贯性和信息性,而另一个单独的奖励则针对最终答案的正确性。对于没有 推理标注的样本,我们仅根据最终答案应用标准奖励。这种双重奖励设计鼓励模型通过初 始化思维和自由思维风格获得一定知识,使模型在医学领域中更好更快地获得有用的知识。

为了促进多样性提示,为了增强模型在不同临床场景中的泛化能力,我们在训练过程中引 入了多样化的提示风格。我们没有使用固定的问答格式,而是让模型接触到一系列反映真 实临床查询的提示类型:

- 需要解释的答案:这些提示鼓励模型在给出答案之前进行彻底的临床推理,答案中应包括相关的理由或支持信息。
- 简短答案:这些提示鼓励提供简洁的事实性或是/否回答,并附以最低限度但在上下 文中相关的理由。
- 开放式答案:这些提示促进自由形式的推理,以处理非结构化或模棱两可的临床情境。

为了进一步研究模型在推理过程中的表现,我们评估了不同提示技术的影响:推理提示和 非推理提示,其定义如下:

- 推理提示:"这是用户和医疗助手之间的一次对话。用户关于图像提出了一个问题,助手解决了这个问题。助手首先一步一步地思考推理过程,然后为用户提供答案。"
- 非推理提示: "用户与医疗助手之间的对话。助手向用户提供答案。"

Table 6: 使用 LLM-as-Judge 评分与 RARL 和提示多样性对医疗 VQA 数据集上的模型进行性能评估。

Dataset	GPT-40 mini		Gemini 1.5 Flash		Human Evaluation		
	Reasoning (%)	Final (%)	Reasoning (%)	Final (%)	Reasoning (%)	Final (%)	
Experiment reasoning prompting in inference							
Ours VQA-RAD SLAKE VQA-Med 2019 Path-VQA	70.27 % 	64.86 % 45.73 % 56.38 % 51.18 % 25.33 %	68.92 % 	64.86 % 44.00 % 54.71 % 50.47 % 24.78 %	70.94 % 	65.54 % 	
Experiment none-reasoning prompting in inference							
Ours VQA-RAD SLAKE VQA-Med 2019 Path-VQA	67.56 	61.49 40.54 % 52.67 % 47.99 % 24.00 %	65.54 % 	61.49 % 40.21 % 52.55 % 46.31 % 23.72 %	65.64 % 	60.81 % 	

类似于第 2.3 节的实验,我们在多个医学 VQA 基准上评估了我们的方法。正如表 6 中总结的那样,将 RARL 与提示多样性相结合显著提升了各数据集的性能。在我们的推理测试集上,与没有 RARL 的情况相比,模型在 LLM 评估和人工评估中提高了大约 4.5 %。与基线相比,由 GPT-4o mini 评估的最终答案分数在 VQA-RAD 上提高了 4.17 %,在 SLAKE 上提高了 9.18 %,在 VQA-Med 2019 上提高了 2.77 %,在 Path-VQA 上提高了 4.41 %。从结果来看,与传统的 RL 方法相比,RARL 和提示多样性的结合显著改善了模型的推理行为。RARL 的双重奖励机制确保模型优先考虑与临床相关的推理,减少了第 2.3 节中观察到的幻觉和通用回复实例。例如,在我们测试集中的推理任务上,模型显示出与真实解释更强的对齐,这通过更高的 LLM 评估分数和人工评估得到了证明。此外,推理和非推理提示之间存在差距。

模型的表现分别由 GPT 4o mini、Gemini1.5 Flash 和人工评估下降了 4.01 %、5.16 % 和 7.78 %。尽管性能有所提升,挑战仍然存在,特别是在推广到分布外数据(Path-VQA 基准)时, 模型的推理能力不够稳健,这表明需要更广泛的训练数据或额外的领域特定微调。

Observation 3 : (1) The experiment shows that reasoning-aware training (RARL) and prompting diversity not only improve performance on reasoning tasks but also help the model internalize more clinically relevant knowledge that generalizes effectively to different types of data. (2) Reasoning prompting during the inference stage is crucial for guiding the model to produce correct answers.

# 3 结论

在这项工作中,我们提出了一个可推理感知的强化学习(RARL)框架,并结合几种策略, 旨在增强医学视觉语言模型(VLM)在数据和硬件资源受限情况下的临床推理能力。通过 使用低秩适应(LoRA)对轻量级模型Qwen2-VL-2B-Instruct进行微调,并采用自定义奖励 函数来共同优化诊断准确性和推理质量,我们的方法在保持计算效率的同时实现了强大的 性能。

我们在一个将大语言模型作为裁判的框架中评估了我们的方法,该框架同时考虑了解释的 正确性和质量。在注重推理的任务中, RARL 比监督微调显著提高了 7.78%,并在未见数据 集上泛化得更好,比 SFT 提高了 27%,比传统强化学习方法提高了 4%。这些结果强调了 在医学视图语言模型中引入推理监督和奖励设计的价值,为更具可解释性和有效性的临床 决策支持系统提供了一条切实可行的路径。此外,我们的研究结果突出了在推理密集的任 务中联合评估推理过程和最终答案的重要性。我们还表明,在推理过程中引导多样性提示 可以补充 RARL,引导小型语言模型朝着临床一致的推理方向发展。

尽管取得了这些进展,挑战仍然存在。模型会出现幻觉或在泛化到分布外数据时遇到困难, 这表明需要更多样的训练数据和强大的领域适应方法。未来的工作将专注于解决这些局限, 通过扩展训练覆盖范围和探索高级提示技术来进一步减少幻觉并改善泛化能力。

## 4

限制

尽管我们提出的流程在有限的数据和硬件限制下能够改善医学 VLM 的推理和泛化,但我们的工作也存在一些局限性,包括数据集、LLM 作为评判者的偏见,以及医学临床医生的评估。

## References

- [1] Tom B Brown. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.
- [2] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.
- [3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023.
- [4] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.
- [5] Qwen Team. Qwen3, April 2025. URL https://qwenlm.github.io/blog/qwen3/.
- [6] Zhihong Huang, Lei Jin, and Hongbo Zhang. Visual-language models for medical image analysis: A survey. IEEE Transactions on Medical Imaging , 42(10):2789–2804, 2023.

- [7] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Machine Learning for Health (ML4H), pages 353–367. PMLR, 2023.
- [8] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, et al. Towards generalist biomedical ai with med-palm m. arXiv preprint arXiv:2402.16994, 2024.
- [9] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915, 2023.
- [10] Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, et al. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv e-prints, pages arXiv–2305, 2023.
- [11] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36:28541–28564, 2023.
- [12] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data, 6(1): 317, 2019.
- [13] Xin Zhang, Chunyuan Li, and Haotian Liu. Pmc-15m: A large-scale dataset for medical visionlanguage pretraining. arXiv preprint arXiv:2305.07890, 2023.
- [14] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. Nature Reviews Cancer, 18(8):500–510, 2018.
- [15] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. arXiv preprint arXiv:2502.19634 , 2025.
- [16] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. arXiv preprint arXiv:2503.13939, 2025.
- [17] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC medical informatics and decision making, 20:1–9, 2020.
- [18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837, 2022.
- [19] Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. Medcot: Medical chain of thought via hierarchical expert. arXiv preprint arXiv:2412.13736, 2024.
- [20] Tan-Hanh Pham, Trong-Duong Bui, Minh Luu Quang, Tan-Huong Pham, Chris Ngo, and Truong-Son Hy. Silvar-med: A speech-driven visual language model for explainable abnormality detection in medical imaging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2025.
- [21] Jun Yao, Ziming Wang, and Wei Chen. Deepseek-rl: Advancing reasoning in large language models with reinforcement learning. arXiv preprint arXiv:2401.08912, 2024.
- [22] Xiaoxiao Chen, Yufan Wang, and Li Zhang. Reinforcement learning for medical image analysis: Current progress and future directions. Medical Image Analysis, 92:103056, 2024.

- [23] Wei Liu, Kai Zhang, and Ming Zhao. Generalized reward-driven policy optimization for visionlanguage models. arXiv preprint arXiv:2403.09876, 2024.
- [24] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280, 2024.
- [25] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. Scientific data, 5 (1):1–10, 2018.
- [26] Bo Liu, Li-Ming Zhan, Li Xu, Lin Qian, Jiahong Wu, and Wenbin Shi. Slake: A semanticallylabeled knowledge-enhanced dataset for medical visual question answering. arXiv preprint arXiv:2102.09542, 2021.
- [27] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In Working Notes of CLEF 2019, volume 2380 of CEUR Workshop Proceedings, Lugano, Switzerland, September 9-12 2019. CEUR-WS.org. URL https: //ceur-ws.org/Vol-2380/paper\_272.pdf.
- [28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.
- [29] Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations , 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.
- [30] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in neural information processing systems, 35:16344–16359, 2022.
- [31] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems, 37:87310–87356, 2024.
- [32] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. arXiv preprint arXiv:2503.05132, 2025.

# A 用于训练视觉语言模型的提示

**Explanation-Required Answer Prompt:** 

```
prompt = f"You are a Medical Assistant. Carefully analyze the medical
image and the User's question. Think through the problem step by
step using clinical reasoning and relevant visual cues. Present your
thinking process clearly within <think> </think> tags, and then provide
a well-justified answer within <answer> </answer> tags. User: question
Assistant: Let me solve this step by step."
Short-Form Answer Prompt:
prompt = f"You are a Medical Assistant. Read the User's question and the
associated medical image. Provide your thinking within <think> </think>
tags and a concise factual answer within <answer> </answer> tags. User:
question Assistant:"
Open-Ended Answer Prompt:
prompt = f"You are a Medical Assistant. The User askes a question based
on the provided medical image. Consider all visual and contextual details.
Share your reasoning freely within <think> </think> tags, and conclude
with your best judgment within <answer> </answer> tags. User: question
Assistant:"
```

# B 作为法官的大型语言模型的提示

```
evaluation_prompt = '''
Task:
```

You are given a question, a ground truth, and a prediction in medical analysis. Evaluate the model prediction for its relevance, accuracy, and alignment with the ground truth.

```
Reasoning Scoring: if the reasoning of the prediction is similar to ground truth,
returning 1, otherwise returning 0.
Prediction score: if the final prediction is similar to ground truth,
returning 1, otherwise returning 0.
Output structure:
{
    "evaluation": Provide a concise justification sentence explaining why you
    rated the score.
    "reasoning_score": reasoning score
    "prediction_score": prediction score
}
;
;
```