影像学报告评估指标及理由和子分数

Yingshu Li[1], Yunyi Liu[1], Lingqiao Liu[2], Lei Wang[3], Luping Zhou[1]*

[1]School of Electrical and Computer Engineering, University of Sydney, NSW 2006, Australia
[2]School of Computer Science, University of Adelaide, SA 5005, Australia
[3]School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia

摘要

自动生成的放射学报告评估仍然是一个基本挑战，因为缺乏临床上可靠、可解释且精细度高的指标。现有方法要么产生粗略的整体评分，要么依赖于不透明的黑盒模型，限制了它们在实际临床工作流程中的应用价值。我们介绍了一个新的放射学报告评估框架**辐射原因**，该框架不仅能输出六个临床上定义的错误类型下的细粒度子分数，还能生成可读的人类解释来说明每个分数背后的理由。我们的方法基于组相对策略优化，并包含两个关键创新：(1) **子分数动态加权**，根据实时 F1 统计数据自适应优先处理临床上具有挑战性的错误类型；以及 (2) **多数指导优势缩放补充模型**，根据子分数一致性和提示难度调整策略梯度更新。这些组件共同使优化更加稳定，并与专家临床判断更好地对齐。在 ReXVal 基准测试上的实验表明，RadReason 超过了所有先前的离线指标，并实现了与基于 GPT-4 的评估相当的结果，同时保持了可解释性、成本效益和适合临床部署的特点。代码将在发表后公开。

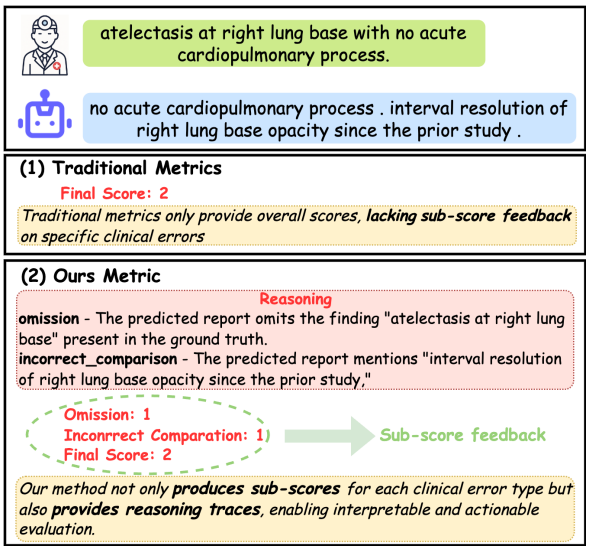

图 1: **不同指标类型评估输出的比较**。传统指标仅提供总体分数，缺乏对特定错误的洞察。我们的方法提供了子分数和推理轨迹，使评估变得可解释且详细。

1 介绍

放射学报告的自动生成功能在临床人工智能中已成为一个关键任务，有望减轻放射科医生的工作负担并提高诊断一致性 (Huang et al., 2023; Li et al., 2023, 2024; Wang et al., 2024) 。然而，评估生成报告的质量仍然是一个根本性的挑战。传统的自然语言生成（NLG）指标，如 BLEU、ROUGE 和 METEOR (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004) ，侧重于词语重叠，未能捕捉临床上有意义的语义差异，尤其是在涉及释义、否定或细微事实错误的情况下。基于嵌入的度量标准，如 BERTScore (Zhang et al., 2019) ，改进了语义对齐，但经常忽视对于临床解释至关重要的领域特定实体。结构感知度量标准包括 RadGraph F1 (Jain et al., 2021) 和 CheXbert F1 (Smit et al., 2020) ，将医学知识纳入考量，但在整体报告质量敏感性和细粒度方面存在不足。RadCliQ (Yu

*Corresponding author

et al., 2023a) 更进一步，通过在多个现有指标上学习回归模型来更好地近似人类判断。最近，大型语言模型（LLMs）已被探索用于放射学报告评估 (Grattafiori et al., 2024; Yang et al., 2024) 。MRScore (Liu et al., 2024) 提出了一种特定于放射学的奖励模型，以实现定制化评分，而 Green (Ostmeier et al., 2024) 则通过显式错误类型匹配来评估事实准确性，实现了与专家判断的高度一致。RaTEScore (Zhao et al., 2024) 使用实体感知相似性改进了语义鲁棒性，能够处理同义词和否定情况。在基于 GPT 的在线应用中，CheXprompt (Zambrano Chaves et al., 2025) 和 FineRadScore (Huang et al., 2024) 利用 GPT-4 识别临床错误类型并通过少量样本提示生成详细更正。

尽管最近取得了进展，现有的评估方法，总结如图 1 所示，面临两个主要局限性：（1）大多数系统仅生成单个总体得分，缺乏错误类型粒度；以及（2）很少有系统为为什么分配特定得分提供明确的理由，从而限制了临床可用性和模型透明度。

为了解决这些问题，我们引入了 RadReason，一种新颖的评估框架，它将报告质量分解为六个临床定义的错误维度（例如，错误预测、遗漏、错误位置）(Yu et al., 2023a)，并为每个生成的报告生成结构化的子分数和相应的自然语言解释。

例如，“该报告未能提及左侧积液 → 遗漏错误 = 1”。

从技术上讲，RadReason 通过分组相对策略优化 (GRPO) (Shao et al., 2024; Guo et al., 2025) 进行训练，这是一种模拟对分组完成偏好的强化学习范式。

然而，与产生单个标量得分的先前工作不同，我们的框架预测了六个不同的子分数，每个子分数对应于特定的错误类型。

这种设置引入了两个主要挑战：（1）某些错误类型很少见或难以处理，需要自适应优先级；以及（2）报告提示的难度各不相同，有些会产生一致的完成结果，而另一些则表现出高度分歧。

为了缓解这些挑战，我们结合了两种辅助机制：（1）子分数动态加权，它根据每个错误类型的 F1 性能动态调整奖励权重，以针对薄弱领域；以及（2）多数引导优势缩放，它利用多数投票统计来估计样本难度并相应地缩放策略梯度更新。

在 ReXVal 基准测试 (Yu et al., 2023b) 上的实验表明，RadReason 实现了与专家评分的最新相关性，优于所有先前的离线指标，同时保持可解释性。

我们的主要贡献包括：

(1) 我们引入了**径向原因**，这是一个基于奖励优化的放射学报告生成评估框架，该框架输出结构化的子分数和自然语言解释。

(2) 我们提出了两种新的训练策略，子分动态加权和多数指导优势缩放，以增强临床敏感性。

(3) RadReason 在 ReXVal 上实现了最先进的与人类一致的性能，同时保持高效、可解释，并且可以扩展到新的评估标准。

2 相关工作

2.1 放射学报告的评估指标。

评估放射学报告生成（RRG）需要能够衡量语言流畅性和临床正确性的指标 (Yu et al., 2023a)。传统的 NLG 指标——如 BLEU (Papineni et al., 2002)，ROUGE (Lin, 2004) 和 METEOR (Banerjee and Lavie, 2005)——主要关注表面层次的重叠，无法捕捉到释义、否定或细微的临床不准确之处。因此，语义上准确但使用了不同措辞的报告可能会受到不公平的惩罚。为了解决语义保真度和临床内容的问题，已经提出了一些领域相关的指标。CheXbert F1 (Smit et al., 2020) 利用一个在 14 种胸部疾病上训练的病理分类模型，而 RadGraph F1 (Jain et al., 2021) 通过结构化的实体关系图来评估事实正确性。RadCliQ (Yu et al., 2023a) 结合多个这样的指标使用与人类标注对齐的回归模型，提高了相关性但解释性有限。最近的研究

利用大型语言模型（LLMs）进行放射学报告评估 (Grattafiori et al., 2024; Yang et al., 2024) 。MRScore (Liu et al., 2024) 训练了一个特定于放射学的奖励模型来定义一个自定义评分框架。Green (Ostmeier et al., 2024) 根据匹配发现和识别错误来评估事实正确性和临床重要性，展示了与专家评估的高度一致性。RaTEScore (Zhao et al., 2024) 是一个实体感知度量，能够稳健地处理同义词和否定，进一步与人类判断保持一致。一些方法还探索了提示商业 LLMs 如 GPT-4。CheXprompt (Zambrano Chaves et al., 2025) 使用 GPT-4 (Achiam et al., 2023) 检测六种特定的错误类型：假阳性、遗漏、不正确的定位、不正确的严重性评级、无关比较和缺失的比较声明。FineRadScore (Huang et al., 2024) 应用少量提示来逐行修正并为每个识别到的错误提供临床严重性评分。然而，这些方法引发了隐私问题，并依赖于在线访问，这限制了它们的实际部署。此外，仍然存在一个共同的问题：大多数指标仅提供单一的整体分数，缺乏细粒度的子方面反馈或得分分配背后的明确理由。为了应对这一挑战，我们开发了一个离线评估框架，能够生成与临床对齐的子评分和每个错误类型的推理解释，从而提升可解释性和实际应用性。

2.2 大型语言模型中的推理。

大型语言模型（LLMs）在模仿人类推理方面表现出强大的能力，特别是通过将复杂任务分解为结构化的中间步骤。这种范式——通常称为显式推理——使模型能够在得出最终输出之前进行可解释的、分步思考。已经提出了一系列遵循此方法的技术，包括基于提示策略如思维链（CoT）(Wei et al., 2022)、面向规划的方法如思维图和思维树 (Besta et al., 2024; Yao et al., 2023) 等。除了提示之外，在带有推理轨迹注释的数据集上进行监督微调（SFT）(Kumar et al., 2025) 可以进一步提升推理能力，但这需要高质量且劳动密集型的标注，限制了可扩展性。为克服这一点，近期的研究采用了强化学习（RL），在没有显式监督的情况下诱导推理行为。例如，DeepSeek-R1 (Guo et al., 2025) 引入了一个 RL 框架，在该框架中模型被引导生成随后产生答案的推理轨迹，并基于最终正确性进行奖励，从而能够从仅包含答案的数据集学习。在此基础上，我们采用了一个针对评估指标定制的 RL 框架。我们的目标是教会模型分配细致的子分并生成自然语言解释。这使得可解释、具体方面的评价成为可能，这对于临床审计和决策支持至关重要。

3 方法

我们提出了一种用于放射学报告评估的强化学习框架，该框架能够生成具有解释性的子分数和跨临床错误类型的解释，如图 2 所示。我们的方法集成了三个关键组件：（1）通过自然语言推理预测子分数；（2）**动态子分数权重**，通过对每个维度的性能进行适应来强调临床上具有挑战性的方面；以及（3）**多数指导优势缩放附加法则**，该组件根据提示难度调节策略更新，以奖励稀有但有价值的完成。

3.1 背景

组相对策略优化（GRPO）(Guo et al., 2025) 是一种用于通过群组偏好信号来优化语言模型的强化学习算法。与成对方法如 DPO (Rafailov et al., 2023) 不同，GRPO 在一个组内比较多个完成情况，并计算相对优势以引导策略更新。对于每个提示，GRPO 采样一组完成情况并通过奖励模型分配奖励。设 $\boldsymbol{r} = \{r_1, r_2, \cdots, r_G\}$ 表示问题 q 的一组 G 完成情况的奖励集合。每个补全的优势在其组内进行归一化：

$$\hat{A}_{i,t} = \tilde{r}_i = \frac{r_i - \text{mean}(\boldsymbol{r})}{\text{std}(\boldsymbol{r})}. \tag{1}$$

GRPO 损失函数可以定义为：

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\left[\frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}\hat{A}_{i,t} - \beta\, D_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\theta_{\text{ref}}}\right)\right], \tag{2}$$

其中，o_i 是第 i 个提示的采样补全体，

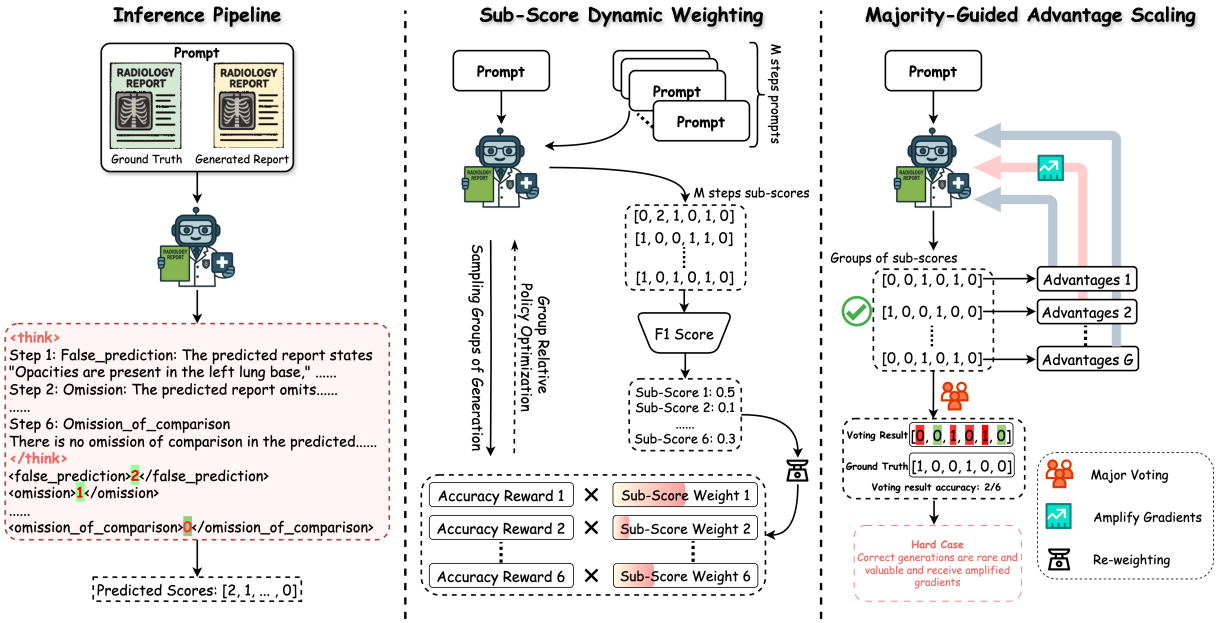

图 2: 概述我们的训练框架。**左：**模型在六个临床定义的错误类型上生成详细的子分数和解释。**中间：**在训练过程中，我们通过定期计算 F1 差距并相应地重新加权每个维度的奖励来应用动态子分数加权。**正确：**我们引入了多数引导优势缩放，它通过比较完成情况中的多数投票子分数与真实值来估计提示难度。我们的方法放大了在困难提示上的正确完成情况的梯度，同时对容易提示上的不正确完成情况进行更重的惩罚。

$\pi_\theta(o_{i,t} \mid q, o_{i,<t})$ 表示当前策略下的词元级可能性。

3.2 奖励构建

为了指导 GRPO 训练，我们设计了三个奖励函数，以捕捉评估质量的不同方面：推理完整性、格式正确性和子分准确性。对于每个采样的完成情况，我们按照以下方式计算奖励：

结构化推理奖励。这个奖励通过验证模型是否明确讨论了在 (Yu et al., 2023a) 中提出的六种临床错误类型来促进可解释性：虚假预测、遗漏、不正确的定位、不正确的严重程度、不正确的比较以及缺少比较。我们使用正则表达式来检测每个方面是否以适当的结构线索被提及（例如，“步骤 1：虚假预测”）。

格式奖励。我们通过要求一个包含推理文本的 <think>...</think>块，并且每个子分标签中恰好有一个有效的数值（例如，<omission>1.0</omission>）来强制输出。这促进了格式的一致性并确保了可靠的子分提取。

准确率奖励。 此组件衡量预测的子分数与真实标注之间的吻合程度。对于每一份生成的报告，我们通过正则表达式提取预测的子分数，并将其与相应的实际值进行比较。与以前基于 GRPO 的方法使用二进制 (0/1) 奖励信号不同，这些方法对任何偏离真实情况的情况都给予零奖励，这样的严格方案无法区分接近正确的预测和完全错误的预测，导致学习信号稀疏且不稳定。为了解决这一问题，我们引入了一个平滑的高斯奖励函数，该函数根据与真实值的距离平方来惩罚预测误差，从而提供更稳定和更有信息量的学习信号。子分数 j 的奖励计算如下：

$$r^{(j)} = \exp\left(-\frac{(\text{pred}^{(j)} - \text{gt}^{(j)})^2}{2\sigma^2}\right), \quad (3)$$

其中 $\sigma = 0.5$ 是控制对预测误差容忍度的标准差。

子分数奖励然后在所有 $K = 6$ 维度上取平均：

$$r_{\text{sub}} = \frac{1}{K}\sum_{j=1}^{K} r^{(j)}. \quad (4)$$

除了单独评估每个子分数外，我们还纳入了一个总分对齐项，鼓励预测的子分数之和与真实总和相匹配：

$$r_{\text{total}} = \exp\left(-\frac{(\hat{s}_{\text{total}} - s_{\text{total}})^2}{2\sigma^2}\right), \quad (5)$$

其中 $\hat{s}_{\text{total}}$ 和 s_{total} 分别表示预测的总分和真实的总分。

最终的准确度奖励是两个组成部分的加权组合：

$$r_{\text{acc}} = r_{\text{sub}} + r_{\text{total}}. \quad (6)$$

单个输出的最终奖励为：

$$r = r_{reasoning} + r_{format} + r_{acc} \quad (7)$$

3.3 子分数动态加权

虽然准确度奖励独立评估每个子分数，但它隐含地假设所有错误类型的重要性相等。实际上，这些方面在频率、模糊性和临床影响上存在差异。例如，省略和错误预测错误更常见但往往具有歧义性，而错误的比较虽然较少见但却经常带有重要的临床意义。因此，统一平均子分数奖励会导致优化偏差，频繁且容易学习的部分会不成比例地影响梯度更新。

为了解决这一不平衡问题，我们提出了一种子分数动态加权 **(SDW)** 策略，该策略能够动态地强调模型表现较弱的子分数。具体来说，每 M 步，我们计算每个方面 $j \in \{1, ..., K\}$ 的 F1 分数 $F1^{(j)}$，并定义相对性能差距为：$\Delta_j = \bar{F1} - F1^{(j)}$，　where　$\bar{F1} = \frac{1}{K}\sum_{j=1}^{K} F1^{(j)}$。

然后我们将这些难度分数转换为一组归一化的权重，使用 softmax 函数：

$$w_j = 1 + \frac{\exp(\alpha \cdot \Delta_j)}{\sum_{k=1}^{K} \exp(\alpha \cdot \Delta_k)}, \quad (8)$$

其中 α 是一个温度超参数，用于控制对表现不佳维度的关注程度。

然后我们计算最终的子分数奖励作为各维度的加权平均值：

$$r_{\text{sub}}^{\text{dyn}} = \frac{1}{K}\sum_{j=1}^{K} w_j \cdot r^{(j)}. \quad (9)$$

这一策略使得模型能够持续地将监督重新分配到临床上具有挑战性或表现不佳的方面。与静态奖励平均不同，SDW 鼓励更加均衡的学习，并提高了在各种错误类型中的鲁棒性。

Algorithm 1 子分数动态加权

Require: F1 分数，更新间隔 M，温度 α
　初始化 $w_j \leftarrow 1$ 为 $j = 1, \ldots, K$
　for each training step $t = 1$ to T **do**
　　if $t \mod M = 0$ **then**
　　计算每个方面的 $F1^{(j)}$ 分数 j
　　计算平均 F1：$\bar{F1} \leftarrow \frac{1}{K}\sum_{j=1}^{K} F1^{(j)}$
　　for each aspect $j = 1$ to K **do**
　　计算 F1 差值：$\Delta_j \leftarrow \bar{F1} - F1^{(j)}$
　　end for
　　使用 softmax 更新权重：

$$w_j \leftarrow 1 + \frac{\exp(\alpha \cdot \Delta_j)}{\sum_{k=1}^{K} \exp(\alpha \cdot \Delta_k)}$$

　　end if
　end for

3.4 多数指导优势缩放依附关系

虽然 GRPO 模型在给定提示的情况下相对衡量了不同补全之间的偏好，但它将所有训练提示视为同等重要，而不考虑它们的难度。然而，并非所有样本都具有相同的学习效用。一些提示较为困难，这体现在其完成质量始终不佳；在这种情况下，高质量生成既罕见又具信息量。相反，其他提示相对容易，在这些提示中大多数生成表现良好，而在此类提示上的错误可能表明模型存在关键故障。

为解决此问题，我们引入了 Majority-Guided Advantage Scaling**(气体管理)** 机制，该机制根据每个提示推断的难度调整优势幅度。具体来说，对于每个提示组，我们汇总所有 G 完成的预测子分数，并对每个 $K = 6$ 子分数维度执行多数表决。对于每个子分数维度，我们在整个组中汇总所有预测值并计算多数表决。然后将此多数表决值与相应的地面实况标签进行比较。如果在定义的一致性阈值下多数预测未能匹配地面实况，则我们认为这是一个困难案例。

我们将所有六个方面的正确性取平均值

来计算一个多数选定的分数 γ。这个分数反映了提示的容易程度，即每个方面上多数预测与真实情况匹配的频率。形式上，设 $P^{(j)} = [p_1^{(j)}, p_2^{(j)}, ..., p_G^{(j)}]$ 表示跨越 G 个完成项的第 j 个子评分方面的预测值集合，并设 $y^{(j)}$ 是真实值。我们定义多数选择的分数 γ 为：

$$\gamma = \frac{1}{K} \sum_{j=1}^{K} \mathbf{1} \left[\text{mode} \left(P^{(j)} \right) = y^{(j)} \right], \quad (10)$$

其中 $\text{mode}(\cdot)$ 返回组中的最频繁值。分数 $\gamma \in [0, 1]$ 反映了多数预测与真实情况相匹配的子分方面比例。缩放方法定义为：

$$s_i(\gamma) = \phi_- + (\phi_+ - \phi_-) \cdot (1 + (\psi_i(\gamma) - c))^{-\beta}. \quad (11)$$

这里，ϕ_- 和 ϕ_+ 分别是缩放的下限和上限。c 是一个难度阈值，而 β 控制着调制的锐度。函数 $\psi(\gamma)$ 定义如下：

$$\psi_i(\gamma) = \begin{cases} \gamma, & \hat{A}_{i,t} > 0 \\ 1 - \gamma, & \hat{A}_{i,t} < 0 \end{cases} \quad (12)$$

最终更新的优势可以定义为：

$$\hat{A}'_{i,t} = s_i(\gamma) \cdot \hat{A}_{i,t}. \quad (13)$$

总结来说，我们介绍了两种策略来增强基于 GRPO 的训练优化。**子分数动态加权**通过根据 F1 表现调整权重，使学习重点放在难以处理的错误类型上。**多数引导优势缩放算法基础**放大了在困难提示下正确完成任务的梯度，并降低了在简单情况下出现错误的权重。这些机制共同引导学习朝向具有临床意义和鲁棒性的行为。

Algorithm 2 主导的优势缩放方法

输入： 预测的子分数组 P，真实值 y，原始优势 $\hat{A}_{i,t}$，缩放参数 (ϕ_-, ϕ_+, c)
输出： 缩放的优势 $\hat{A}'_{i,t}$
for each aspect $j = 1, \ldots, K$ **do**
  收集预测 $P^{(j)} = [p_1^{(j)}, p_2^{(j)}, ..., p_G^{(j)}]$
  计算多数预测：$m^{(j)} \leftarrow \text{mode}(P^{(j)})$
  $\gamma^{(j)} \leftarrow \mathbf{1}[m^{(j)} = y^{(j)}]$
end for
计算一致性：$\gamma \leftarrow \frac{1}{K} \sum_{j=1}^{K} \gamma^{(j)}$
for each completion $i = 1, \ldots, G$ **do**
  if $\hat{A}_{i,t} > 0$ **then**
    $\psi_i \leftarrow \gamma$
  else
    $\psi_i \leftarrow 1 - \gamma$
  end if
  计算缩放因子：
  $s_i \leftarrow \phi_- + (\phi_+ - \phi_-) \cdot (1 + (\psi_i(\gamma) - c))^{-\beta}$.
  更新优势：
  $\hat{A}'_{i,t} \leftarrow s_i \cdot \hat{A}_{i,t}$
end for

4 实验与结果

4.1 数据集

训练数据生成。 我们从 MIMIC-CXR 数据集中抽样了 1,000 份放射学报告作为基准参照。对于每个案例，我们提示 GPT-4 生成具有不同错误特征的合成诊断报告。之前的研究显示 GPT-4 在评估任务中与专家放射科医生的一致性很高 (Liu et al., 2024)。通过系统地注入临床错误（例如，遗漏、虚假发现、定位错误）来控制报告质量，从而生成不同精确度的报告。具体而言，指示 GPT-4 生成：(1) 高质量报告包含 0 – 1 个错误；(2) 中等质量报告包含 2 – 3 个错误；(3) 低质量报告包含 4 个或更多错误。

此基于错误计数的提示功能能够对语义保真度进行精细控制，确保在整个临床可能的质量水平范围内进行全面覆盖。总共我们收集了 3,968 个带有标签的报告完成情况，每个都与其对应的地面实况锚点配对。详细提示见附录 A。

重 X 值 (Yu et al., 2023b) 是一个公开的基准，旨在评估自动指标与放射报告评估中专家人类判断之间的对齐情况。它由来自 50 个 MIMIC-CXR 研究中的 200 对候选-参考报告组成，每个研究包含四个候选报告。每一对都由六位具有董事会认证的放射科医生使用六个类别的 RadCliQ 错误分类法进行标注，该分类法基于每个类别中的错误数量提供详细的子分数标注。

4.2 实验设置

我们评估了所提出的指标与一系列已建立的基线，包括传统的 NLG 指标（BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005)），语义相似度指标（BERTScore (Zhang et al., 2019)）以及临床导向或结构感知的指标如 RadGraph F1 (Jain et al., 2021)，Semb 分数和 RadCliQ-v1 (Yu et al., 2023a)。遵循先前的工作 (Yu et al., 2023b)，我们通过计算 ReXVal 基准上指标输出与专家注释之间的 Kendall's Tau 和 Spearman 相关性来评估与人类判断的一致性。

我们采用 Qwen2.5-7B[1] (Team, 2024) 作为基础语言模型。使用 LoRA (Hu et al., 2022) 进行微调，秩为 16，缩放因子为 $\alpha = 16$。我们将学习率设置为 1e−5，批量大小设为 4，并使用 2 个 ×NVIDIA RTX A6000 (48GB) GPU 训练了 2 轮。对于多数指导优势缩放，我们使用 $\phi_- = 0.8$、$\phi_+ = 1.2$ 和难度阈值 $c = 0.5$。

4.3 主要结果

与现有指标的比较。表 1 展示了我们提出的指标与 ReXVal 数据集上一组全面的基线方法进行比较评估的结果，使用 Kendall' s Tau 和 Spearman 相关系数来衡量与专家评分的一致性。我们的方法在所有非大语言模型的方法中取得了最高的整体性能，Kendall' s Tau 为 0.730，Spearman 相关系数为 0.871。它始终优于传统的 NLG 指标（BLEU-4、ROUGE-L、METEOR）、语义相似度分数（BERTScore、Semb Score）以及结构感知的临床指标（RadGraph F1、RadCliQ-v1）。最近先进的模型，如 GREEN (Kendall: 0.640) 和 RaTEScore (0.527)，缺乏细粒度的可解释性且不提供子分数，在详细评估场景中的实用性受到限制。在线方法如 CheXprompt 和 FineRadScore 达到了相当的 Kendall 分数（分别为 0.750 和 0.737），但依赖于商业大语言模型 API（例如 GPT-4），这引入了与隐私、可重复性和运营成本相关的问题。

[1]https://huggingface.co/Qwen/Qwen2.5-7B

4.4 消融研究

为了评估每个组件的贡献，我们在表 2 中总结了一个消融研究。从一个通过监督微调训练的基础模型开始，我们逐步引入了 GRPO、子分数动态加权（SDW）和多数指导优势缩放（MGAS）。

引入 GRPO 显著提高了与专家标注的一致性（肯德尔：0.495 → 0.690；斯皮尔曼：0.634 → 0.832），证实了基于偏好的学习在评估放射学报告中的有效性。我们部分归因于 GRPO 在生成过程中更好地捕捉中间推理信号的能力。添加 SDW 在遗漏的发现和缺乏比较等子评分上进一步改进，突显了其在动态强调临床上代表性不足但关键的错误类型方面的实用性。最后，结合 MGAS 带来了整体最佳性能（肯德尔：**0.730**，斯皮尔曼：**0.871**）。这验证了我们的直觉，即正确回答困难提示时应增加奖励信号放大，而如果在容易的提示上出现错误则应更加严厉地惩罚。MGAS 通过将学习信号与案例难度对齐来帮助稳定更新。

总体而言，这些结果展示了 SDW 和 MGAS 在提升 GRPO 优化动力学方面的互补优势，实现了在各种误差类别中的高度相关性。

4.5 定性分析

图 3 提供了在 RadCliQ 框架下，我们模型的子分数预测如何与人类标注对齐的视觉示例。在第一种情况下，生成的报告错误地引入了一个新的发现 "左侧少量胸腔积液"，并遗漏了关键短语 "左上肺叶切除术后改变"（两者都已突出显示），导致 "虚假预测" 和 "漏报" 的分数均为 1.0，完全与专家评分一致。在第二种情况下，模型虚构了 "肺血管充血" 和 "右侧少量胸腔积液"，同时遗漏了临床上重要的改善情况 "肺水肿" 和 "右上叶密度增高"。这些错误导致 "虚假预测"、"漏报" 和 "不正确的严重程度" 的分数升高，与人类评估结果非常接近。这些示例展示了模型执行细粒度错误识别的能力，这与专家判断一致。

Ground truth report: post left upper lobectomy changes, with no superimposed acute intrathoracic process detected.

Predicted report: small left pleural effusion. no superimposed acute intrathoracic process.

Reasoning	Ours	Human
Step 1: false_prediction - The predicted report mentions "small left pleural effusion" which is not mentioned in the ground truth. This is an additional finding not present in the ground truth report.	**1.0**	**1.0**
Step 2: omission - The ground truth mentions "post left upper lobectomy changes" which is omitted in the predicted report.	**1.0**	**1.0**
Step 3: incorrect_location - The predicted report does not misidentify the location of any finding.	0.0	0.0
Step 4: incorrect_severity - The predicted report does not misassess the severity of any finding.	0.0	0.0
Step 5: incorrect_comparison - No comparison is made in either report.	0.0	0.0
Step 6: omission_of_comparison - No comparison is mentioned in either report.	0.0	0.0

Ground truth report: ap chest compared to ___: previous mild but asymmetric pulmonary edema continues to improve. the residual opacification in the right upper lobe raises concern for pneumonia. heart size is normal. there is no pleural effusion.

Predicted report: compared to chest radiographs ___ . pulmonary vascular congestion which developed between ___ and ___ persists . very mild interstitial pulmonary edema has improved , but small right pleural effusion has accumulated . the absence of the vasculature in the right upper lobe is usually an indication of emphysema . heart size is normal .

Reasoning	Ours	Human
Step 1: false_prediction - The predicted report mentions "pulmonary vascular congestion" and "small right pleural effusion" which are not mentioned in the ground truth. It also incorrectly states that the absence of vasculature in the right upper lobe indicates emphysema instead of pneumonia.	**2.0**	**1.667**
Step 2: omission - The predicted report omits the continuation of improvement in asymmetric pulmonary edema and the residual opacification in the right upper lobe that raises concern for pneumonia.	**1.0**	**0.667**
Step 3: incorrect_location - Not applicable in this case as the locations of the findings are not misidentified.	0.0	0.0
Step 4: incorrect_severity - The predicted report downgrades the severity of pulmonary edema from "mild but asymmetric" to "very mild."	**1.0**	**0.667**
Step 5: incorrect_comparison - Not applicable in this case.	0.0	0.333
Step 6: omission_of_comparison - Not applicable in this case.	0.0	0.333

图 3: 案例研究比较我们模型的子分数预测与人类标注，每种六类临床错误类型均展示了逐步推理。颜色高亮标记推理结果。

5 结论

我们介绍**径向原因**，一个用于放射学报告的可解释评估框架，该框架生成结构化的子分数和跨临床意义的错误类别的明确推理。通过嵌入子分数动态加权和多数指导优势缩放附加条件，我们的方法可以自适应地关注更难的子方面，并根据提示难度校准学习。在 ReXVal 基准测试上的经验结果表明，RadReason 不仅优于先前的指标。

限制

我们的评估是在 ReXVal (Yu et al., 2023b) 上进行的，这是与人类判断一致的标准放射学报告评估基准。尽管由于专家标注的成本而规模适中，但它能够实现方法之间的公平和有意义的比较。我们采用了 RadCliQ (Yu et al., 2023a) 中定义的六个基于临床的错误类别；虽然固定，但我们的框架是模块化的且易于扩展到其他或分层的分类法。尽管我们的实验集中在来自 MIMIC-CXR 的胸部 X 光报告上，但奖励驱动的方法是对模态无偏见的，并可以推广到结构化诊断输出，如 CT、MRI 或多模式报告中。展望未来，基于子分数的奖励公式也可能启发其他临床生成任务（例如医疗 VQA）的评估方法。

度量	肯德尔的 Tau↑	斯皮尔曼 ↑
BLEU-4 (Papineni et al., 2002)	0.345	0.475
ROUGE-L (Lin, 2004)	0.491	0.663
METEOR (Banerjee and Lavie, 2005)	0.464	0.627
BertScore (Zhang et al., 2019)	0.507	0.677
RadGraphF1 (Jain et al., 2021)	0.516	0.702
Semb_score (Yu et al., 2023a)	0.494	0.665
RadCliQ-v1 (Yu et al., 2023a)	0.631	0.816
GREEN (Ostmeier et al., 2024)	0.640	–
RaTEScore (Zhao et al., 2024)	0.527	–
我们的	**0.730**	**0.871**
以下结果并非严格可比		
因为他们正在使用一个在线模型（例如，*GPT4*）。		
RadFact (Bannur et al., 2024)	0.590	–
CheXprompt (Zambrano Chaves et al., 2025)	0.750	–
FineRadScore (Huang et al., 2024)	0.737	–

表 1: 人类相关性比较在 ReXVal 数据集上的评估指标。

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. 2024. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. In *Machine Learning for Healthcare Conference*. PMLR.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan.

准则	基线		+ 组群		+ 软件定义网络		+ MGAS（我们的）	
	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman	Kendall	Spearman
False prediction	0.507	0.581	0.608	0.704	0.614	0.704	**0.645**	**0.742**
Omission of finding	0.323	0.366	0.576	0.662	0.583	0.682	**0.596**	**0.706**
Incorrect location	0.375	0.401	0.461	0.482	**0.533**	**0.571**	0.473	0.506
Incorrect severity	0.430	0.460	0.571	0.614	0.450	0.482	**0.570**	**0.611**
Absence of comparison	0.106	0.112	0.170	0.182	0.176	0.189	**0.186**	**0.196**
Omission of comparison	0.160	0.168	0.194	0.204	**0.317**	**0.333**	0.238	0.252
总计	0.495	0.634	0.690	0.832	0.698	0.838	**0.730**	**0.871**

表 2: 奖励设计的消融研究。我们的完整模型结合了 GRPO 训练、子分动态加权（SDW）和多数指导优势缩放（MGAS），实现了最高的相关性，用粗体显示。

2025. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*.

Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2023. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11.

Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. 2024. Kargen: Knowledge-enhanced automated radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 382–392. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yunyi Liu, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2024. Mrscore: Evaluating medical report with llm-based reward system. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 283–292. Springer.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Md, Michael Moseley, Curtis Langlotz, Akshay Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, Chuanfu Li, and Jin Tang. 2024. Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. *arXiv preprint arXiv:2410.00379*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan.

2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023a. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, EKU Fonseca, Henrique Lee, Zahra Shakeri, Andrew Ng, and 1 others. 2023b. Radiology report expert evaluation (rexval) dataset.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, and 1 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019.

A 附录

A.1 GPT-4 提示模板

图 4 中的提示用于生成 GPT-4 的训练数据。

GPT-4 PROMPT

You are a professional radiologist. Your task is: given a ground truth diagnostic report, generate three predicted reports with systematically injected clinical errors to simulate varying levels of report fidelity. These predicted reports should be rated based on the following error-counting rules. Specifically, the predicted reports should fall into three fidelity levels:

1. High-quality: 0–1 errors
2. Medium-quality: 2–3 errors
3. Low-quality: 4 or more errors

Please control generated report quality by intentionally injecting semantic-level clinical errors, including:

1) false_prediction (False report of a finding in the predicted report), "
2) omission (Missing a finding present in the ground truth), "
3) incorrect_location (Misidentification of a finding's anatomic location/position), "
4) incorrect_severity (Misassessment of the severity of a finding), "
5) incorrect_comparison (Mentioning a comparison that isn't in the ground truth), "
6) omission_of_comparison (Omitting a comparison detailing a change from a prior study). "

Please generate three predicted reports for the given ground truth report between <ground_truth_report><\ground_truth_report>. Then, for each predicted report, list its errors by category.

<ground_truth_report>YOUR GROUND TRUTH REPORT<\ground_truth_report>

图 4: GPT-4 提示 示例。