arXiv:2305.03660v1 [cs.CL] 5 May 2023

使用 OpenAI GPT 模型的检索增强型胸部 X 射线报告生成

Mercy Ranjit[*] MERANJIT@MICROSOFT.COM
印度蒂鲁吉拉伯利的巴拉蒂达桑大学计算机科学

Gopinath Ganapathy GGANAPATHY@GMAIL.COM
印度蒂鲁吉拉伯利的巴拉蒂达桑大学计算机科学系

Ranjit Manuel[†] RANJIT.F.MANUEL@GMAIL.COM
用于数字健康与影像的人工智能，
印度班加罗尔印度科学研究所

Tanuja Ganu TAGANU@MICROSOFT.COM
Microsoft Research，
印度班加罗尔

编辑： 编辑姓名

摘要

我们提出将检索增强生成（RAG）作为一种自动放射学报告撰写方法：利用对比预训练视觉语言模型的多模态对齐嵌入，为输入的放射学图像检索相关的候选放射学文本；再利用 OpenAI text-davinci-003、gpt-3.5-turbo 和 gpt-4 等通用领域生成模型，根据检索到的相关放射学文本生成报告。该方法能够控制幻觉生成，并利用这些生成模型的指令遵循能力，按照我们期望的格式生成报告内容。我们的方法取得了更好的临床指标，BERTScore 达到 0.2865（$\Delta+$ 25.88 %），S_{emb} score 达到 0.4026（$\Delta+$ 6.31 %）。我们的方法广泛适用于不同的临床场景，因为它既能在自动放射学报告生成过程中加入与特定场景相关的内容，也能在报告生成过程中将用户意图和要求注入提示词，从而根据临床场景调整生成报告的内容与格式。

1. 引言

自动放射学报告生成系统可以通过多种方式改进放射科医生撰写报告的工作流程。这些人工智能系统可以生成自由文本内容或结构化报告内容，供放射科医生审核各种感兴趣的属性，例如器官。

[*] MR 也隶属于班加罗尔的 Microsoft Research
[†] RM 也隶属于班加罗尔的 Databricks Inc.

© M. Ranjit、G. Ganapathy、R. Manuel 和 T.Ganu。

系统、病理、异常、严重程度、影像发现的大小或位置、进展状态等。

一些关于 AI 辅助放射学报告生成的现有研究，将放射学报告生成问题视为图像描述生成问题或生成任务 [Chen 等人(2020); Miura 等人 (2020)]。近期一项有趣的研究 CXR-RePaiR [Endo 等人(2021)] 利用与放射学图像相关的有限诊断细节和属性，将其视为一个检索问题。这种方法的优势在于，它能够利用庞大的既往和当前放射学报告数据库，同时生成影像印象建议。不过，所有这些方法都存在各种问题，例如生成内容中包含无关内容或产生幻觉。

Ramesh 等人 提出的最新研究 CXR-ReDonE (2022年)针对报告生成中与既往影像参考相关的问题之一，通过创建一个新数据集 CXR-PRO [Ramesh 等人] ，移除了 MIMIC-CXR 放射学文本报告中的既往影像参考 [Johnson 等人 (2019)] ，并使用该数据集改进 CXR-RePaiR [Endo 等人(2021)] ，从而在放射学报告印象生成领域建立了新的最先进基准。

但是，像 CXR-RepaiR 和 CXR-ReDonE 这样的纯基于检索的报告撰写模型，不允许用户根据自身需求调整检索结果，而这对于在各种临床场景中的适用性很有帮助。以下是纯基于检索的方法的一些局限性，其中一些在这些论文中被列为未来工作：

- 纯基于检索的报告生成可能会在无异常发现病例报告中包含无关的检索结果，尤其是在为生成报告而设置相关检索结果数量 K 时。 > 1

- 生成内容中还可能出现一些不必要的噪声，例如既往报告引用、提及医生姓名、用户特定细节等；如果存在，这些内容就会被原样提取。纯基于检索的报告生成也会受到检索结果中内容重复的影响。

- 此外，也无法按特定格式生成放射学报告，而不同的报告生成下游应用可能会有此要求。与其输出原始文本，不如提取病理、位置信息、严重程度、大小等感兴趣的属性，以生成结构化放射学报告，这可能会更有帮助。

- 纯检索系统生成的报告中可能包含一些不连贯的信息，因为检索结果可能会原样带来属于两位不同患者的报告中的句子。

目前已有能力非常强大的生成式大语言模型（LLM），例如 text-davinci-003、gpt-3.5-turbo（Chat GPT）和 gpt-4，可用于通用领域，并能在零样本或少样本设置下，根据指令式提示为各种下游任务生成相关内容。探索如何利用这些模型协助放射科医生完成放射学报告生成工作，将大有裨益。然而，这些模型缺乏医学领域场景所特需的最新信息或领域专属信息。在生成时为这些模型提供最新且相关的领域专属内容，有助于扩展大语言模型的能力，使其能够处理训练阶段未接触过的数据任务。

此外，我们还可以利用这些模型遵循指令的能力，引导它们生成我们所需的回答。为大语言模型提供生成内容所需的额外上下文，可以减少其产生幻觉的情况。我们受到 Lewis 等人 (2020) 的研究中所验证的检索增强生成（RAG）优势的启发。该研究表明，RAG 生成的内容能够更充分地以真实事实知识为依据，从而减少幻觉；论文的“更广泛的影响”部分还指出了其在各种下游任务中的广泛应用。

我们提出了基于对比式X射线-报告配对检索的生成（CXR-RePaiR- Gen），这是一种用于放射学报告生成的基于检索增强生成（RAG）的方法，在 CXR-RepaiR 的工作基础上进行了扩展 [Endo 等人 (2021)] 和 CXR-ReDonE [Ramesh 等人 (2022年)]。如 图1所示，我们利用经过对比式预训练的 ALBEF 模型 [Li 等人 (2021)]，为放射学报告数据库生成视觉-语言对齐嵌入。随后使用同一模型为输入的放射学图像生成嵌入。由于图像嵌入和文本嵌入在对比式预训练过程中已完成对齐，因此会根据输入图像嵌入与放射学报告嵌入之间的相似度，为输入的 X射线图像检索最相关的放射学文本（报告或句子）。最后，利用 OpenAI 的 text-davinci-003、gpt-3.5-turbo 和 gpt-4 模型，从筛选后的记录集中生成一份整合的放射学报告印象。

基于检索增强生成的方法不仅使放射学报告生成能够以从放射学文本语料库中检索到的相关放射学文本为依据，还允许用户通过提示工程将用户意图作为指令、将少样本示例作为生成过程的一部分，以生成符合临床环境要求格式的内容。

关于医疗保健领域机器学习的可泛化见解

我们的方法为医疗保健中的机器学习带来了以下关键见解：

- 检索增强生成能够结合各种特定领域医疗保健领域编码器模型与通用领域生成模型的优势，充分发挥两类模型的长处。我们证明，这可以改善临床指标。

- 我们还通过将大语言模型生成的响应与从放射学报告或句子语料库中检索到的放射学文本进行比较，评估放射学报告生成中的幻觉。这有助于在规划于真实临床环境中实际使用这些系统时做出决策。

- 我们的论文还展示了如何利用大语言模型中的提示工程，通过少样本学习注入用户意图和需求，从而生成适用于下游应用的不同输出格式的放射学报告。

我们的方法凭借 BERTScore [Yu 等人 (2022年)] 达到 0.2865（$\Delta+$ 25.88%），并且 S_{emb} 得分 [Endo 等人 (2021)] 达到 0.4026（$\Delta+$ 6.31%），优于此前最先进的检索方法 CXR-ReDonE。在临床场景中，这些得分的提升意味着我们能够生成在语义上更接近真实印象的放射学报告，同时报告非常简洁，减少了来自

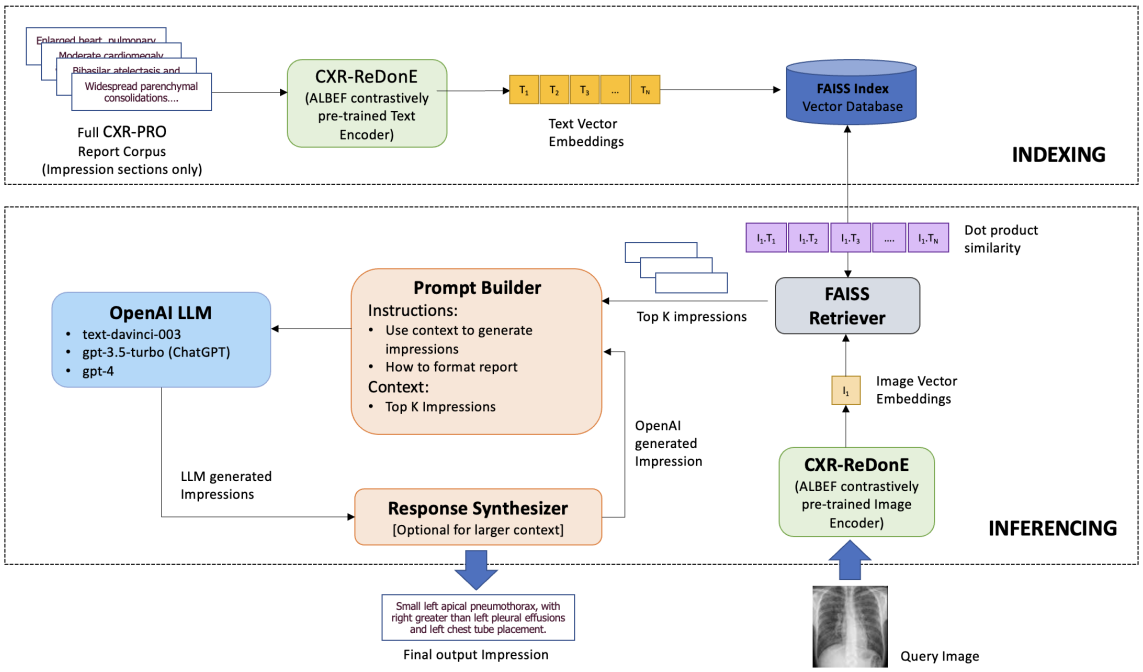

图1：我们使用经过对比预训练的视觉语言编码器（CXR-ReDonE），将放射学印象中各句子的所有文本嵌入投影到向量数据库索引中，并使用同一编码器模型，根据输入图像嵌入检索最匹配的句子。检索到的影像印象报告或句子与生成印象的指令一起，构成提供给大语言模型的提示词上下文。

检索结果。我们在 RadGraph 上也与 CXR-ReDonE 持平 F_1 指标 [Yu 等人(2022年)]。该指标衡量我们是否能够准确检索所有临床实体。由于检索增强生成基于从 CXR-ReDonE 检索模型获得的检索记录，因此在这一指标上我们无法超越 CXR-ReDonE。

2. 相关工作

近期放射学报告生成研究将该问题视为一种生成任务，例如 Chen 等人（2020）使用 Transformer 解码器架构 R2Gen 的工作，以及 Miura 等人（2020）的工作。后者采用一种名为 M2 Trans 的基于奖励的强化学习方法，专注于生成完整、一致且临床准确的报告。

Endo 等人 (2021在其研究 CXR-RePaiR 中，使用检索方法解决放射学报告生成问题，并采用更具临床可靠性的指标建立了新的最先进基准。检索基于他们使用 MIMIC-CXR 数据集进行对比式预训练的视觉语言模型 [Johnson 等人 (2019)]。论文中引入了一种新的临床疗效相似度指标 S_{emb}，用于利用 CheXbert [Smit 等人 (2020)]标注器的最后隐藏表示计算语义相似度

参考报告和预测报告。论文还使用了 BERTScore 指标[Zhang 等人(2019)] 作为衡量语义相似度的另一项指标。

Ramesh 等人（2022年）在其研究中解决了所有自动放射学报告生成方法都面临的一个关键问题：放射学报告中的既往报告引用会影响报告生成质量。他们通过解决 MIMIC-CXR 数据集中的这一问题，构建了一个新的 CXR-PRO 数据集[Ramesh 等人] [Johnson 等人（2019）]。他们还使用 CXR-PRO 数据集和更新后的 ALBEF 架构重新训练了 CXR-RepaiR [Li 等人（2021）] ，并在放射学报告生成任务上创下当前最先进水平。他们还使用 RadGraph F1 [Yu 等人（2022年）] 分数作为额外指标，借助 RadGraph 模型衡量预测报告中临床实体的完整性和准确性 [Jain 等人（2021）]

随着大语言模型的兴起，Lewis 等人（2020）在其研究中引入了检索增强生成（RAG），其关键优势之一是利用外部知识源来补充大语言模型的知识，从而完成任务。大语言模型生成的内容也有坚实的真实知识基础，因此“幻觉”较少，生成结果也更符合事实。论文的广泛影响声明提到，该方法可应用于各种场景，例如为其配备医学索引。

在本工作中，我们为大语言模型提供放射学报告文本的索引，并将其用作知识库，使大语言模型能够针对输入的放射学图像生成放射学报告印象。为了实现图像和文本的多模态检索，我们使用了来自 CXR-ReDonE [Ramesh 等人 （2022年）] 的对 CXR-RePaiR 工作进行改进的 [Endo 等人 （2021）] 对比预训练视觉语言模型，以研究在这些检索结果之上进行增强生成是否能够进一步提升报告生成基准。我们还研究能否利用大语言模型的通用能力，根据用户要求调节报告生成输出，从而支持不同临床场景中的多种使用模式。我们还评估检索增强生成产生的幻觉，以便将所提出方法的应用转向真实世界的临床环境。

3. 方法

本文提出将放射学报告生成视为一项数据增强生成任务，并使用 `text-davinci-003`、`gpt-3.5-turbo` 和 `gpt-4`等大语言模型。我们的假设是，无需使用领域专用生成模型，而是需要领域专用检索器。我们可以利用领域专用编码器生成的嵌入来执行数据检索任务，并使用通用领域生成模型对检索到的数据进行增强。我们在 CXR-RePaiR 的工作基础上 [Endo 等人 (2021)] 以及 CXR-ReDonE[Ramesh 等人 (2022年)] 展开，将放射学报告生成问题视为检索任务，并添加一个利用检索到的印象进行生成的步骤。我们使用 CXR-ReDonE 中的对比预训练视觉语言模型 ALBEF [Li 等人(2021)]，从放射学图像和报告中生成图像与文本嵌入。为选择用于数据增强的前 K 条记录，我们尝试了报告级语料库 $R = r_1, \ldots\ldots, r_n$ 以及句子级语料库 $S = s_1, \ldots\ldots, s_n$，与 CXR-RePaiR 中的做法相同。

我们使用 ALBEF [Li 等人 (2021)] 模型（来自 CXR-ReDonE），利用 CXR-PRO 数据集为放射学报告文本语料库生成具有对比对齐特性的文本嵌入

[Ramesh 等人] ，并将其编入向量数据库。我们使用相同的模型为输入放射学图像 x 生成对比对齐图像嵌入，并根据点积相似度，使用该嵌入从报告或句子语料库中检索前 K 条记录。选取与输入图像嵌入相似度最高的前 K 份报告或句子，用于增强使用 `text-davinci-003`、`gpt-3.5-turbo` 和 `gpt-4` 模型进行的生成。我们通过提示词 P 向大语言模型提供提示词，并将从句子语料库 S 中检索到的前 K 个句子作为上下文，同时附上用于生成的指令 Q，从而生成印象 I。

$$I = LLM(P(Q, \sum_{i=1}^{k} S_i))$$

其中，使用函数 $argmax_{s \in S(R),}\, f(s, x)$ 从 S 中选取前 K 个句子，其中 f 表示输入句子 s 与放射学图像 x 之间的相似度点积函数，与 CXR-RePaiR 中的定义一致。

如果用于大语言模型生成的上下文记录大小超出了大语言模型的令牌限制，我们建议采用精炼方法，通过大语言模型迭代生成响应。我们首先将从放射学文本语料库中检索到的前 K 条记录中的第一份报告或第一个句子作为上下文，通过提示词输入大语言模型，从而生成初始印象。随后，我们将大语言模型生成的初始印象响应与第二个句子或报告作为上下文，再依此类推，最终从大语言模型获得经过精炼的答案。 I_k 表示使用精炼提示词迭代构建的大语言模型印象，P^r 表示利用前几轮迭代生成的印象和指令 Q 构建的提示词，S_k表示句子语料库中的第 k 个句子。

$$I_k = LLM(P^r(Q, I_{k-1}, S_k))$$

Langchain 等基于大语言模型的框架提供了精炼机制 [Chase (2022年)]以及 LlamaIndex [Liu (2022年)]。由于检索增强生成中使用的上下文大小可能因不同的临床场景而异，本文提出采用精炼方法。在本研究的实验中，由于我们使用了较小的 $K = 1,2,3$ 个值，且上下文仍在大语言模型的限制范围内，因此无需采用精炼机制，也不需要通过迭代方式逐步构建响应。

除了生成自由文本放射学报告之外，我们还假设，将放射学报告以结构化格式呈现会更加有用，其中包括检索结果中的感兴趣的属性。这些感兴趣的属性可以是病理、与病理相关的严重程度、严重程度、大小或位置等。我们使用提示工程和少样本示例来生成结构化放射学报告输出。

3.1. 检索语料库

我们以 CXR-PRO 数据集的训练集印象部分构建检索语料库 [Ramesh 等人] 该数据集包含 374,139 份自由文本放射学报告及其对应的胸部 X 线影像。由于 CXR-PRO 基于经过去标识化处理的 MIMIC-CXR 数据集，因此不包含任何受保护健康信息 (PHI)。CXR-PRO 是 MIMIC-CXR 数据集的一个改编版本 [Johnson 等人(2019)]]，其中省略了既往影像参考。它解决了

放射学报告生成模型会产生幻觉式既往影像参考这一问题。我们使用语料库中放射学报告的印象部分，并将报告级印象以及构成报告级印象的句子均视为用于报告生成的检索语料库，这与 CXR-RePaiR 中的做法一致 [Endo 等人(2021)]

3.2. 基线方法

我们将 CXR-ReDonE [Ramesh 等人 （2022年）] 提出的基于检索的报告生成方法（使用 CXR-PRO 数据集） [Ramesh 等人] 作为基线方法。我们旨在探究，在这些检索结果的基础上使用大语言模型进行检索增强生成，是否有助于改善放射学报告生成的临床指标。

3.3. 提示设计

我们设计了两组提示词，用于将放射学报告生成为自由文本报告：一组用于 `text-davinci-003` 模型，另一组用于在对话设置下使用 `gpt-3.5-turbo` 和 `gpt-4` 模型进行报告生成，如表 1所示。我们向大语言模型提供指令，要求其使用检索到的句子作为上下文来生成放射学报告。对于`gpt-3.5-turbo` 和 `gpt-4` ，提示设计包括对话设置中的系统提示和用户提示。系统提示指示系统根据用户将发送的上下文生成放射学报告的印象。用户提示将检索记录作为上下文发送，并请求系统以放射学报告作为响应。

3.4. 结构化报告输出

我们还通过在提示词中的少样本示例里指定期望的报告输出格式，尝试调控报告生成的输出。对放射科医生和下游应用而言，除了生成自由文本放射学印象之外，从放射学报告中提取某些感兴趣的属性也可能很有用。这些感兴趣的属性包括提取病理、与病理相关的严重程度、严重程度、影像发现的大小或位置等。我们指示大语言模型以结构化输出格式生成放射学报告，其中包含印象摘要以及检索上下文中出现的感兴趣的属性。我们会说明感兴趣的病理及其他属性，并提供少样本提示词，如表 2中提示设计所示。

3.5. 实验

我们使用基于 CXR-ReDonE 嵌入的检索记录，采用 OpenAI 大语言模型 text-davinci-003、gpt-3.5-turbo 和 gpt-4 进行了检索增强生成实验。实验中，我们同时考虑了基于报告的语料库和基于句子的语料库。语料库中的检索记录构成提示词中的上下文，大语言模型据此生成自由文本放射学印象。我们针对自由文本印象生成采用了零样本设置，并针对结构化报告生成采用了少样本设置。

表 1：在零样本设置下，将检索到的报告作为上下文文本，用于 OpenAI 大语言模型进行放射学报告印象生成的提示词。提示词中的斜体文本对应于用于格式化提示词的变量。

text-davinci-003	**gpt-3.5-turbo/gpt-4 系统提示**	**gpt-3.5-turbo/gpt-4 用户提示**
生成 an 印象 摘要 的放射学 报告，并使用给定的上下文。 严格遵循指示 以下 在生成 印象时。 说明： • 印象 摘要 应 be 基于 on 其中的信息 上下文。 • 将生成内容限制为 最多 个词。 上下文: 最大长度 印象摘要：	你是一名专门设计的助手。 撰写印象总结 用于放射学报告。 用户将发送一段上下文文本。 您将收到并回复一条 印象 摘要 使用 该上下文。 说明： • 印象 应 be 基于用户提供的信息，用户将在 上下文中发送。 •印象应 not 提及任何内容 关于后续行动。 •印象应 not 包含任何提及AV在线 既往或以往研究。 • 将生成内容限制为 最大长度词数。	上下文： 上下文 印象摘要：

表 2：在少样本设置下，将检索到的报告作为上下文文本，为 OpenAI 大语言模型设计的提示词，用于生成结构化放射学报告。提示词中方括号内的文本对应于格式化提示词时使用的变量。

提示设计	**少样本示例**

为放射学报告生成印象摘要，所用的上下文。
印象中的病理应来自列表词语，例如：{病理}
位置词应从词语列表中选取例如：{位置词}。
严重程度词应从词语列表中选取例如：{严重程度词}
大小词应从词语列表中选取，例如：{大小词语}
上下文：{上下文示例}
印象：{报告 JSON 示例}。。
上下文：{上下文示例}
印象：{报告 JSON 示例}。。
上下文：{上下文示例}
印象：{示例报告 JSON}。。
上下文：{上下文}
印象：
	上下文：
	右侧肺门上方阴影 may 与.....有关肺部 血管 充血 尽管感染过程或误吸不能完全排除，需结合适当的临床环境。
	
	印象：
	{
	”印象”：“轻度双侧肺底肺不张是存在。右侧肺门上方阴影可能是由于肺部血管充血。”,
	”属性”：[
	{
	”病理”: ”肺不张”,
	”位置”: ”双侧肺底”
	},
	{
	”病理”: ”阴影”,
	”位置”：”右侧肺门上区”
	}]
	
	}

4. 真实数据上的结果

4.1. 评估数据集

我们在两个黄金基准数据集上评估性能：MS-CXR [Boecking 等人]，以及 CXR-PRO 数据集中的测试集印象 [Ramesh 等人] 。这两个数据集都是在经委员会认证的放射科医生协助下创建的。CXR-PRO 包含 2,188 张放射学图像及其相关报告。CXR-PRO 数据集经过预处理，以删除重复行。MS-CXR 数据集包含 1,162 个图像–句子对，涵盖八种不同的心肺放射学影像发现。尽管这两个数据集来自相同的数据分布，但它们并不相同：MS-CXR 是一个短语定位数据集，还包含用于在放射学图像中定位短语的边界框，其短语非常精确、简洁；而 CXR-PRO 的印象文本则较长。我们在这两种设置下评估性能，因为在某些临床场景中，可能需要针对放射学图像给出精确、简洁的短语。

4.2. 评估方法

我们使用从报告级语料库和句子级语料库的检索记录中生成的大语言模型放射学印象自由文本来进行评估。对于句子级语料库，我们评估基于前 K 个句子检索结果生成的印象，其中 K= 1、2、3。我们的基线是从 CXR-ReDonE 中检索到的印象。我们采用两个语义指标进行评估——BERTScore [Zhang 等人 （2019）] 和 S_{emb} [Endo 等人 （2021）]，以考察生成结果与真实印象之间的相似度。在医学语境中，这种评估更具意义，因为诸如肺萎陷这样的短语可以表示肺不张，即使句子中没有出现完全相同的词语。BERTScore 会计算预测印象中的每个词元与真实印象中的每个词元之间的相似度分数。词元级相似度使用上下文嵌入进行计算，而不是直接匹配词元。 S_{emb}使用 CheXbert 模型 [Smit 等人 （2020）] 来计算最终隐藏状态表示中嵌入之间的余弦相似度。为了评估生成报告和真实报告中所包含临床实体的重叠情况，我们使用 RadGraph F_1，这是一项由 Yu 等人 （2022年）提出的指标，使用 RadGraph 模型 [Jain 等人（2021）]来评估临床实体的重叠情况。

我们还以结构化 JSON 格式生成放射科报告印象，以评估能否根据用户要求格式化生成的印象输出。目前尚未对其进行定量测量，但我们展示了结构化生成的定性输出。

4.3. 幻觉评估

我们还定性和定量地评估报告印象摘要是否会根据提供给大语言模型作为上下文的前 K 个句子的语料库产生幻觉。我们使用大语言模型生成结果与前 K 个上下文记录之间的 S_{emb} score，判断它们在语义上是否等价。

表 3：在 CXR-PRO 测试印象上评估 CXR-RePaiR-Gen（基于检索增强生成的方法），用于报告语料库检索和句子语料库检索。评估指标包括 BERTScore、S_{emb} score 和 RadGraph F_1。对于所有 K 值，在基于报告和句子语料库的检索中，我们的方法在临床指标 BERTScore 和 S_{emb} score 上均优于基线；在 k=3 时，RadGraph F_1 指标与 CXR-ReDonE 持平。斜体表示优于基线，粗体表示我们的方法在 RadGraph F_1 最高值处取得的最高分，该值对应 k=3。需要指出的是，基于检索增强生成的 CXR-RePaiR-Gen 在 RadGraph F_1 score 上不能超过 CXR-ReDonE，否则就意味着我们对上下文中不存在的临床实体产生了幻觉。

K	方法	评估指标		
		BERTScore	Semb	RadGraph F1
N/A	CXR-ReDonE	0.2482	0.3647	0.0921
	CXR-RePaiR-Gen（text-davinci-003）	*0.2600*	*0.3741*	0.0839
1	CXR-ReDonE	0.2455	0.4029	0.0861
	CXR-RePaiR-Gen (text-davinci-003)	*0.2610*	*0.4116*	0.0774
2	CXR-ReDonE	0.2465	0.3892	0.1045
	CXR-RePaiR-Gen (text-davinci-003)	*0.2753*	*0.4036*	0.0926
3	CXR-ReDonE	0.2276	0.3787	0.1104
	CXR-RePaiR-Gen(text-davinci-003)	*0.2782*	*0.4030*	0.1018
	CXR-RePaiR-Gen（gpt-3.5-turbo）	*0.2748*	*0.3973*	0.0991
	CXR-RePaiR-Gen(gpt-4)	***0.2865***	***0.4026***	**0.1061**

4.4. 结果 – CXR-PRO

我们以 CXR-ReDonE 中纯检索式方法生成的放射科报告印象作为基线来评估性能。我们发现，基于检索增强生成的生成方式提升了报告和句子语料库检索的 BERTScore 指标，在基于句子的检索中，k=3 时绝对提升了 5.06%。类似地，它还提升了 S_{emb} 报告和句子语料库检索的得分，在基于句子的检索中，k=3 时绝对提升了 2.43%。RadGraph F_1这一衡量临床实体检索情况的指标，在 k=3时与 CXR-ReDonE 持平，在较低的 k 值时略低。需要指出的是，我们的方法基于增强后的数据生成印象，因此 RadGraph F_1 指标不可能超过 CXR-ReDonE，否则就意味着我们生成了不以上下文为依据的临床实体。评估指标见表 3。

4.5. 结果 – MS-CXR

我们的方法将 MS-CXR 短语定位文本上的 BERTScore 绝对提升了 8.67，并且 S_{emb} 绝对提升了 3.86。评估指标可见于表 4。

表 4：使用 text-davinci-003 进行基于句子语料库的检索（k $= 3$），评估 CXR-RePaiR-Gen 在 MS-CXR 短语上的表现。评估指标为 BERTScore、 S_{emb} 和 Rad- Graph F_1。我们的方法在临床指标 BERTScore 和 S_{emb} 上均优于基线方法，在 RadGraph F_1方面则表现相当。斜体表示优于基线方法的结果，粗体表示我们的方法取得的最高值。

方法	**评估指标**		
	BERTScore	**Semb**	**RadGraph F1**
CXR-ReDonE	0.1102	0.3494	0.0626
CXR-RePaiR-Gen(text-davinci-003)	***0.1970***	***0.3880***	**0.0617**

4.6. 定性分析

与纯检索策略的输出相比，我们发现基于检索增强生成的印象非常简洁、冗余信息更少，同时仍保留了所有相关临床实体。请参阅表 5 和表 6，查看基于检索增强生成的方法分别为 CXR-PRO 和 MS-CXR 中的示例生成的简洁印象摘要。检索增强生成也会尽量避免检索语料中无关紧要的细节。在表 5的最后一个样本中，基于检索增强生成的输出会自动排除随访建议等细节，从而使印象更加简洁。

表 5：CXR-PRO 数据集示例：生成的报告与参考报告及基线方法 CXR-RedonE 的比较。基于检索增强生成的 CXR-RepaiR-Gen 生成结果提供了非常简洁的印象摘要，同时能够保留重要的临床实体。输出是在检索结果的 Top K 值设为 3 时生成的。

真实标注	**CXR-ReDonE (K=3)**	**CXR-RepaiR-Gen (K=3) (text-davinci-003)**	**CXR-RePaiR-Gen (K=3) (gpt-4)**

右肺门肿块 与右中叶不张相关，并且邻近有明显的淋巴管癌病和实变。右肺门肿块与不张 相关，累及右中叶，并且邻近有明显的淋巴管癌病 和实变。

虽然这可能反映了致密的实变性肺炎，但影像发现至少令人担忧其可能存在潜在肿块，伴随肺炎症也可能是另一种诊断。右中叶融合性实变区域，以及右侧肺门上区，可能代表继发感染或肺不张。右中叶实变部分由已知的潜在转移性疾病所致，也可能继发阻塞后感染或肺不张。

大型右肺门周围肿块伴有广泛肺不张以及右中叶融合性实变区域和右侧肺门上区，令人担忧其可能存在潜在恶性肿瘤，肺炎症也可能是另一种诊断。

疑似右肺门肿块，令人担忧存在阻塞性淋巴结肿大。右中叶融合性实变及肺门上区实变，可能由感染或肺不张所致。影像发现提示右肺门周围存在大型肿块，伴有广泛肺不张，提示可能为恶性肿瘤。

正位及侧位胸片：肺容积明显偏低，使轻度心脏增大显得更为明显；但肺血管充血，并伴有轻度间质性水肿，见于下肺，提示心功能失代偿。

清除不佳。右上肺中度肺水肿，残余水肿可能是右肺底残余混浊影的原因；除非后续证实其对利尿治疗无反应。轻度肺水肿，伴中度心脏增大、少量右侧胸腔积液及血管上肺野分布。

轻度肺水肿，伴中度心脏增大、少量双侧胸腔积液及血管上肺野分布，反映心功能失代偿。

轻度肺水肿，中度心脏扩大，右侧少量胸腔积液，血管影上移，右侧肺门周围阴影可能存在感染。

胸部前后位：肺部两侧胸腔呈朦胧样密度增高容积减小，使心影显得严重伴有双侧胸腔积液增大，且伴有压迫性肺不张。心影及肺底部阴影增大，心影中度增大，伴有双侧胸腔积液中等双侧胸腔积液心影增大。可见及肺血管增大，淤血。下肺野心影增大，伴有肺部朦胧样密度增高肺血管提示淤血，双侧胸腔积液呈分层状阴影，肺底部有一些不张改变，与显著分层积液相符，且压迫性双侧阴影肺底不张。符合胸腔积液及肺底部压迫性肺不张。	双侧胸腔积液，压迫性肺不张，心影增大，且可见肺血管淤血。不能排除轻度肺水肿和间质性水肿。	双侧模糊阴影，胸腔积液分层，压迫性肺不张，心影增大，肺血管淤血、双肺底阴影，以及可能存在的轻度肺水肿。

4.7. 指令驱动的输出

我们还通过提供特定提示词，要求排除生成的印象中对既往报告的引用以及后续行动的任何提及，从而评估 GPT-4 遵循指令的能力。

CXR-PRO 评估结果：

- 检索上下文文档中与随访建议相关的 40 处内容均未纳入最终印象。
- 在 87 条提及既往报告的内容中，有 80 条提及成功从印象中排除。

根据提示词中的指示，移除了既往报告引用和后续操作的一些案例见表7

4.8. 调节报告生成格式

检索增强生成的一项关键优势是能够通过提示工程调节输出生成格式。表 8 展示了使用表2中的少样本提示提取感兴趣属性后，以结构化 JSON 格式生成的报告。

表 6：MS-CXR 数据集示例：生成的报告与参考报告及基线方法 CXR-RedonE 的比较。基于检索增强生成的 CXR-RepaiR-Gen 生成结果避免了冗余和重复细节，同时保留了重要的临床实体。输出是在 top K 检索结果设为 3 时生成的。

真实标注	**CXR-ReDonE (K=3)**	**CXR-RepaiR-Gen（K=3）(text-davinci-003)**
左侧肺门周围轻微密度增高肺门区	**模糊的不对称阴影-左下叶的异常** **哪个 可能代表一个 area 早期的 感染。** **左侧肺门下区密度影模糊-** **sity ，** 可能 相关联的伴有下呼吸道炎症的-感染或感染，或潜在-可能为早期或轻度支气管肺炎肺炎 ，但尚无明确临床意义。**性质不确定的 3 cm 卵圆形密度影**位于左肺 基底部伴有 可能的心影后对应影见于侧位片。	不明确的不对称密度影，位于左下叶提示早期感染或正在发展的肺炎，具体取决于临床情况。
左下叶实变由肺炎引起	基底部阴影更为融合异常可能反映重力依赖性水肿 ， but 误吸 and 感染性 肺炎 应该 also be 考虑 in the 合适 临床环境。心后区 阴影可能 反映 肺底肺不张- 胸部影像数据库 ，然而感染或误吸 may 生成 a 相似 外观 ， **and** **短期随访影像检查可能会有所帮助** **如有临床需要，可考虑进行** **。右侧 内部颈静脉导管尖端位于右心房 and is needs 被拉回去** **表示唇缘交界或上方 ，3. 5 厘米必须是向后拉。**	轻度肺水肿伴有斑片状阴影和双侧的积液， 主要地 on 左侧。 压迫性肺不张-不张和肺炎呈阳性-可能。

表7：CXR-PRO 数据集中随机抽取的案例示例，其中基于 RAG 的 gpt-4 生成结果根据基于指令的提示词，消除了对既往报告和随访建议的引用。

上下文记录	**RAG印象**
鉴别诊断 表现包括间质性 lung 疾病 with or 未见 叠加 新发肺炎与 伴随右侧膈穹隆抬高 膈肌。斑片状右肺底部阴影可能反映 压迫性肺不张，但感染 不能排除，且**后续复查** **利尿后建议进行图表检查**。 双肺底部模糊阴影，右侧多于左侧 左侧，考虑到肺容量较低，可能是由于肺不张所致 体积，然而无法排除潜在感染 在适当的临床环境中应予以排除。	间质性肺病，可能出现 肺炎或肺不张，右侧斑片状 肺底部阴影，以及双肺底部模糊阴影， 右侧较左侧明显。 不能排除 感染。
单侧左侧胸腔积 液**并不常见，尤其是患者此前曾有** **右侧胸腔积液时；应考虑** **心包炎、胰腺炎或左上** **腹部象限病变等原因**。 **获得**。 双侧阴影伴轮廓征 显示半侧膈肌轮廓消失，符合分层性 胸腔积液及其下方压迫性 肺底不张。双侧可见分层性积 液，伴斑片状肺泡腔阴影 主要位于肺底部，左侧多于 右侧，可能代表斑片状肺不张。	单侧左侧积液，双侧阴影 伴膈肌轮廓消失、压迫性肺底不张， 以及斑片状肺泡腔阴影 主要位于 肺底部， left 大于右侧。

表 8：使用表 2 中的少样本提示词，通过 CXR-RepaiR-Gen 生成的结构化报告结果表 2。我们可以看到，能够从检索上下文中以结构化格式生成印象摘要，并提取病理、位置信息、严重程度和大小。

结构化报告输出（JSON）

{
”印象”:”可见 Swan-Ganz 导管尖端位于右肺动脉近端。
Dobbhoff 管位置恰当，延伸至胃内。 双侧重度
下叶肺不张及少量至中量胸腔积液”

”影像发现”:[

{

”病理”:”肺不张”,
”位置”:”双侧，基底部”,
”严重程度”:”严重”,
”大小”:””
},

{
”病理”:”胸腔积液”,
”位置”:”双侧”,
”严重程度”:””,
”大小”:” small to moderate”
}

] }

4.9. 检索增强生成中的幻觉

我们计算了 S_{emb} 生成的报告印象与使用 K=3 个检索结果从句子语料库中检索到的检索上下文之间的分数。我们发现，平均相似度得分为 0.8466，约 87% 的测试集印象具有 $S_{emb} > 0.70$ ，这表明生成结果没有出现幻觉。表9展示了测试集中几个 S_{emb} 得分最低的记录。我们发现，即使是这些记录也没有出现幻觉，而最低得分可能归因于生成的印象摘要较为简洁。

5. 讨论

我们发现，CXR-RepaiR-Gen——一种基于检索增强生成的方法，使用对比预训练模型 CXR-ReDonE 的嵌入来检索相关放射学文本，并使用 OpenAI 大语言模型进行生成，可帮助生成非常简洁、准确的放射科报告印象，并提升临床疗效指标，尤其是 BERTScore 和 S_{emb} 分数，同时保留相关临床实体。我们还发现，通过少样本提示和指令进行基于大语言模型的报告生成，可以按需调整报告输出的内容和格式，有助于减少噪杂文本（例如进一步评估的建议、既往报告提及内容、具体患者详情等），并避免在印象中重复内容；这些都是纯粹基于检索的设置中常见的问题。我们注意到，CXR-RePaiR论文提到了这些局限性，因此有必要使用模板数据库去除这些噪杂细节或重复内容，并利用该数据库执行检索，以生成质量更高的报告。我们发现，基于检索增强生成的方法可通过有效的提示工程来解决这一局限性，也可采用少样本提示下的结构化内容提取等方法，仅提取感兴趣的属性，并根据检索记录生成更简洁、准确且完整的印象摘要。

5.1. 局限性

我们注意到，由于RAG印象生成的效果对从语料库中检索到的句子较为敏感，而语料库基于对比预训练模型生成的嵌入，因此，除非允许RAG使用其已有的通用知识，否则RAG生成的临床实体将受到检索结果中临床实体的限制。在实验中，我们通过提示词指令限制模型使用上下文进行生成，以减轻幻觉。因此，检索模型必须能够为生成过程带来所有相关的临床实体。

5.2. 未来工作

我们发现，基于RAG的报告生成可以受益于更先进的对比模型，这些模型能够更加敏感地捕捉放射学图像中的细节，例如严重程度、大小、位置、病理以及放射学图像中的其他感兴趣的属性。医学文本提示工程的进展也是一个值得探索的领域，从而能够更高效地引导大语言模型完成特定的下游任务。

表9：检索到的上下文句子语料库与大语言模型生成结果之间 Semb分数较低的案例示例。Semb 使用 CheXbert 模型计算最终隐藏状态表示所产生的嵌入之间的余弦相似度，并可用于判断是否存在幻觉。我们发现不存在幻觉。

S_{emb}	**上下文记录 (K=3)**	**RAG印象**
0.1508	右肺底部肺容量- 肺不张与实变相比- dation which could represent 右侧临床背景下的肺炎 。肺容积低 伴有双肺底部阴影， 可能代表肺不张，而非 误吸。 肺容积低- 双侧肺底积聚，伴有双肺底部阴影 这可能是由于 为肺不张。	右肺底不张和 实变 可能 due to 肺炎。 双侧肺底阴影- 可能是由肺不张引起的。 或误吸。 肺容量低- 存在体积。
0.2116	左肺总体仍 清晰，但存在极轻微的 左侧线状阴影 肋膈角处，可能 代表炎症后瘢痕 或亚段性肺不张。 不张。 有迹象表明 左肺尖有微小的气胸 ，这可能意味着此前曾 尝试从左侧置管 ，之后才从右侧置管。 可见.....增大 心影，提示 有轻度充血 左侧肺血管 ，因此不排除 肺部不对称升高的可能 肺静脉压。	左肺清晰， 最小线性 阴影 at 左侧肋膈角。 左肺尖微量气胸- 胸部及轻度充血的 左侧肺血管 提示不对称升高 肺静脉压。 心影增大- 注明的注射器。

表 10：使用句子语料库进行基于检索增强生成的报告生成时，text-davinci-003 模型在不同温度设置下的评估指标，top K 检索结果设为 K=3。

温度	评估指标		
	BERTScore	Semb	RadGraph F1
T=0	0.2782	0.4030	0.1018
T=1	0.2783	0.3996	0.1013
T=2	0.2787	0.3998	0.1011

6. 附录

6.1. GPT-3 超参数

我们使用 text-davinci-003 模型进行零样本影像印象生成，并尝试了 0、0.5 和 1 三种温度值，同时从句子级语料库中检索排名前 K 的检索结果（K=3），并将其作为上下文。我们发现，正如 表 10 所示，这些温度值对输出分数的影响并不显著。要求将检索到的句子语料库用作生成上下文的指令可能是导致这一现象的原因。

参考文献

Benedikt Boecking、Naoto Usuyama、Shruthi Bannur、Daniel C Castro、Anton Schwaighofer、Stephanie Hyland、Maria Wetscherek、Tristan Naumann、Aditya Nori、Javier Alvarez- Valle 等人。MS-CXR：充分利用文本语义以改进生物医学视觉-语言处理。

Harriso n Chase。Langchain，2022年。网址 https://github.com/hwchase17/langchain .

ZhihongChen、Yan Song、Tsung-Hui Chang 和 Xiang Wan。通过记忆驱动型 Transformer 生成放射学报告。*arXiv* 预印本 *arXiv:2010.16056*，2020 年。

Mark Endo、Rayan Krishnan、Viswesh Krishna、Andrew Y Ng 和 Pranav Rajpurkar。使用预训练对比语言-图像模型的基于检索的胸部 X 射线报告生成。载于健康机器学习，第 209–219 页。PMLR，2021 年。

Saahil Jain、Ashwin Agrawal、Adriel Saporta、Steven QH Truong、Du Nguyen Duong、Tan Bui、Pierre Chambon、Yuhao Zhang、Matthew P Lungren、Andrew Y Ng 等人。《Rad-graph：从放射学报告中提取临床实体和关系》。*arXiv* 预印本 *arXiv:2106.14463*，2021 年。

Alistair EW Johnson、Tom J Pollard、Nathaniel R Greenbaum、Matthew P Lungren、Chih- ying Deng、Yifan Peng、Zhiyong Lu、Roger GMark、Seth J Berkowitz 和 Steven Horng。Mimic-cxr-jpg，一个大型公开的已标注胸部 X 线影像数据库。*arXiv* 预印本 *arXiv:1901.07042*，2019。

Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich K¨uttler、Mike Lewis、Wen-tau Yih、Tim Rockt¨aschel 等人。《检索增强

生成，用于知识密集型 NLP 任务。神经信息处理系统进展，33：9459–9474，2020。

Junnan Li、Ramprasaath Selvaraju、Akhilesh Gotmare、Shafiq Joty、Caiming Xiong 和 Steven Chu Hong Hoi。《先对齐再融合：结合动量蒸馏的视觉与语言表征学习》。神经信息处理系统进展，34：9694–9705，2021 年。

Jerry Liu。LlamaIndex，2022年。网址 https://github.com/jerryjliu/gpt_inde x.

Yasuhide Miura、Yuhao Zhang、Emily Bao Tsai、Curtis P Langlotz 和 Dan Jurafsky。改进图像到文本放射学报告生成的事实完整性和一致性。*arXiv* 预印本 *arXiv:2010.10042*，2020 年。

Vignav Ramesh、Nathan Chi 和 Pranav Rajpurkar。Cxr-pro：省略既往参考内容的 MIMIC-CXR。

Vignav Ramesh、Nathan A Chi 和 Pranav Rajpurkar。通过移除对不存在的既往影像的幻觉性参考，改进放射学报告生成系统。载于健康机器学习，第 456–473 页。PMLR，2022年。

Akshay Smit、Saahil Jain、Pranav Rajpurkar、Anuj Pareek、Andrew Y Ng 和 Matthew P Lungren。Chexbert：结合自动标签器和专家标注，使用 BERT 准确标注放射学报告。*arXiv* 预印本 *arXiv:2004.09167*，2020。

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、AndrewY Ng 等。评估自动生成胸部 X 光放射学报告的进展。*medRxiv*，页码 2022–08，2022年。

Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。Bertscore：使用 BERT 评估文本生成。*arXiv* 预印本 *arXiv:1904.09675*，2019。