·论著·

DOI：10.3969/j.issn.0253-9802.2024.11.001

生成式人工智能在生成影像学报告方面的表现评估

欢迎扫码观看
文章视频简介

黎超，陈优美，段亚妮，陈耀萍，陈秀珍，覃杰✉
（中山大学附属第三医院放射科，广东 广州 510630）

【摘要】 **目的** 评估 2 种生成式人工智能（AI）在生成腹部影像学报告方面的表现，并与人类医师进行比较。**方法** 回顾性研究 2023 年 6 月至 2024 年 5 月在中山大学附属第三医院接受腹部 CT 和 MRI 检查的 300 例患者的影像学报告。使用生成式 AI 模型 ERNIE 4.0 和 Claude 3.5 Sonnet 对 300 例患者的影像学所见重新生成影像学报告，由 5 名放射科医师采用五点 Likert 量表（1 表示强烈不同意，5 表示强烈同意）评估其完整性、准确性、表达、幻觉和无修改接受度。采用 Friedman 和 Nemenyi 检验进行统计学分析。比较生成式 AI 与人类医师的表现差异。**结果** 研究共纳入 300 例患者的影像学报告。在完整性方面，Claude 3.5 Sonnet 与人类医师相当，均优于 ERNIE 4.0［（4.86 ± 0.37）分 *vs.*（4.76 ± 0.46）分 *vs.*（4.40 ± 0.64）分，前两者比较 $P = 0.200$，前两者与后者比较 P 均 < 0.01］。在准确性方面，人类医师优于 2 种 AI 模型［（4.96 ± 0.22）分 *vs.*（4.66 ± 0.57）分 *vs.*（4.69 ± 0.57）分，前者与后两者比较 P 均 < 0.01］。在无修改可接受度方面，Claude 3.5 Sonnet 与人类医师相当，均优于 ERNIE 4.0［（4.64 ± 0.53）分 *vs.*（4.69 ± 0.54）分 *vs.*（4.30 ± 0.59）分，前两者比较 $P = 0.595$，前两者与后者比较 P 均 < 0.01］。在表达和幻觉上，三者比较差异无统计学意义（P 均 > 0.05）。**结论** Claude 3.5 Sonnet 生成的影像学报告与人类医师水平相当。这提示先进的生成式 AI 有潜力辅助人类医师的工作，有助于提高效率并减轻认知负担。

【关键词】 生成式人工智能；自然语言处理；影像学报告；腹部

Evaluation of the performance of generative artificial intelligence in generating radiology reports

LI Chao, CHEN Youmei, DUAN Yani, CHEN Yaoping, CHEN Xiuzhen, QIN Jie✉
(*Department of Radiology, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China*)
Corresponding author: QIN Jie, E-mail: qinjie@mail.sysu.edu.cn

【**Abstract**】 **Objective** To evaluate the performance of two categories of generative artificial intelligence (AI) in generating abdominal radiology reports, and compare with the performance of radiologists. **Methods** The radiology reports of 300 patients who underwent abdominal CT scan and MRI in the Third Affiliated Hospital of Sun Yat-sen University from June 2023 to May 2024 were retrospectively studied. The generative AI models of ERNIE 4.0 and Claude 3.5 Sonnet were utilized to re-generate radiology reports of 300 patients. Five radiologists evaluated the comprehensiveness, accuracy, expressiveness, hallucinations, and acceptance without revision of the impressions using a five-point Likert scale. Friedman test and Nemenyi test were used to compare the performance between two models and radiologists. **Results** CT and MRI reports from 300 patients were evaluated. For comprehensiveness, Claude 3.5 Sonnet was on a par with human physicians, and both were superior to ERNIE 4.0 (scores of 4.86 ± 0.37 *vs.* 4.76 ± 0.46 *vs.* 4.40 ± 0.64; comparison between the first two, $P = 0.200$, comparison between the first two and the third, both $P < 0.01$). For accuracy, Radiologists outperformed both ERNIE 4.0 and Claude 3.5 Sonnet (scores of 4.96 ± 0.22 *vs.* 4.66 ± 0.57 *vs.* 4.69 ± 0.57; comparison between the first and the latter two, both $P < 0.01$). For acceptance without revision, Claude 3.5 Sonnet was on a par with human physicians, and both were superior to ERNIE 4.0 (scores of 4.64 ± 0.53 *vs.* 4.69 ± 0.54 *vs.* 4.30 ± 0.59; comparison between the first two, $P = 0.595$, comparison between the first two and the third, both $P < 0.01$). Expressiveness and hallucinations metrics showed minimal variations among the three (all $P > 0.05$). **Conclusions** Claude 3.5 Sonnet yields comparable performance to radiologists in generating radiology reports, indicating that advanced generative AI has the potential to assist radiologists, improve the work efficiency and reduce cognitive burden.

收稿日期：2024-10-13
基金项目：国家自然科学基金青年科学基金项目（82202129）；广东省自然科学基金面上项目（2017A030313841）；中山大学附属第三医院国家自然科学基金培育专项项目（2021GZRPYM06）；中山大学附属第三医院“五个五”工程项目（2023WW605）
作者简介：黎超，主治医师，研究方向：人工智能在医学影像学中的应用，E-mail: lich356@mail.sysu.edu.cn；覃杰，通信作者，主任医师，研究方向：心胸影像学，E-mail: qinjie@mail.sysu.edu.cn

【Key words】 Generative artificial intelligence; Natural language processing; Radiology report; Abdomen

目前，患者影像学检查需求激增程度已经远远超过了放射科医师增长数量[1-3]，这种失衡导致工作负荷增加、诊断发布延迟，且增高了医疗人员的职业倦怠风险[4]，影响了医疗服务效率与诊疗护理质量[5]。近年来，人工智能（AI）在医学中的应用越来越广泛[6-7]。其中，生成式 AI 技术凭借其快速生成详细文本的能力，为影像学报告生成自动化提供了一种潜在的解决方案，有助于提高放射科医师的工作效率并减轻工作负担[8-10]，使他们能够专注于复杂病例和关键决策上[11-12]。

近期研究探索了生成式 AI 在影像学报告生成方面的潜力，但结果不一，引发了争议。Sun 等[13]评估了 GPT-4 在生成 50 份胸部 X 线报告方面的表现，认为 AI 生成的结论不如放射科医师。然而这一结论受到了 Ray[14] 的质疑，他们强调了 AI 的潜力和进一步研究的必要性。其他研究显示 AI 在辅助医学文本生成方面有积极的成果，例如用于影像学报告的迭代优化框架提示了放射科医师与 AI 协作改善报告质量的好处[15]，而使用预训练变换器自动生成放射学报告的初步评估为 AI 在这一领域的潜力提供了早期证据[16]。尽管取得了这些进展，但现有研究仍受限于样本量小、缺乏放射学专家的全面评估以及简化的提示可能无法充分利用 AI 的能力。此外，一些研究过度依赖自动化指标而非放射科医师的评估[15, 17]，可能忽视了临床实用性和可解释性的关键方面[18-19]。

本研究评估了生成式 AI ERNIE 4.0 和 Claude 3.5 Sonnet 在影像学报告生成方面的表现，通过扩大样本量并将研究范围扩大到更复杂的腹部 CT 和 MRI 检查以解决前期研究的局限性。并且在采用了模拟放射科医师诊断过程的高级提示工程技术的基础上，进一步由专业放射科医师对模型进行评估。这些改进有助于更全面地探讨生成式 AI 在影像学报告生成方面的能力。现将研究结果报告如下，以期为同行们进一步合理应用生成式 AI 生成影像学报告提供参考。

1 对象与方法

1.1 研究对象

本研究回顾性收集 2023 年 6 月至 2024 年 5 月在我院进行腹部 CT 或 MRI 检查的患者的影像学报告。初步纳入该期间的所有腹部 CT 和 MRI 报告（共 43 648 份）。在排除了影像所见部分少于 100 字的报告后，从剩余报告中随机选取 300 例独立患者的 300 份报告进行分析（筛选报告时仅选择“检查部位”字段含有“上腹部”“下腹部”或“盆腔”的病例，涵盖广泛的疾病类型，不限于典型特征）。详细的数据选择过程代码已公开发表在 GitHub 平台（https://github.com/lichao312214129/code_for_impressionGeneration）。随机数据选择程序在’random_data_selection.py’脚本中实现。除上述 300 份报告外，本研究还纳入了 2021 年 12 月的 35 份 CT 和 MRI 报告，专门用于提示工程，以指导生成式 AI 基于影像学检查结果生成结论。在每份报告中，所有受检对象的个人健康信息和潜在可识别数据均被删除。研究设计见图 1。本研究获得我院伦理委员会的批准（批件号：中大附三

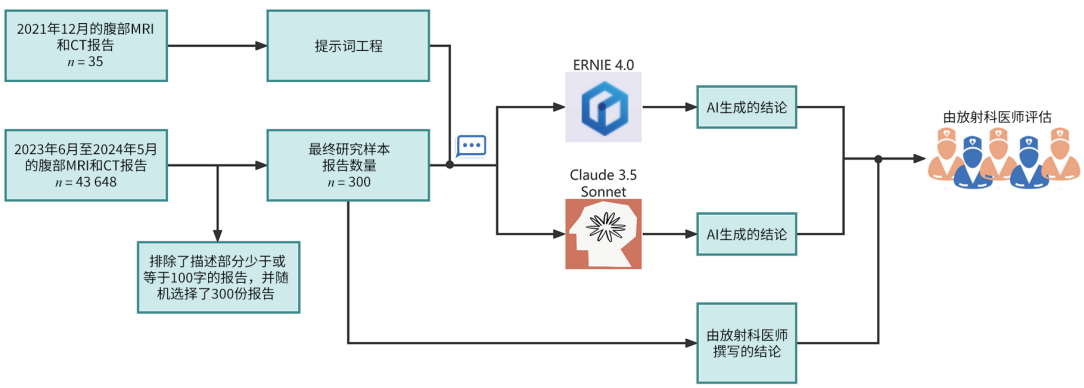


图 1 评估 AI 和放射科医师生成的影像学报告研究设计流程图

Figure 1 Flow chart of study design for evaluating AI and radiologist generated imaging reports

医伦Ⅱ 2023-042-01），并因使用非身份识别数据而豁免了知情同意。

1.2 生成式 AI 的选择

于 2024 年 6 月 13 日至 7 月 5 日期间访问 ERNIE 4.0（ERNIE-4.0-8K-Latest，https：//qianfan.cloud.baidu.com）和 Claude 3.5 Sonnet（claude-3-5-sonnet-20240620，http：//claude.ai）。选用这 2 个模型是基于它们在生成式 AI 领域的领先地位和广泛认可度。ERNIE 4.0 在本研究进行时被认为是中国先进的大型语言模型之一；Claude 3.5 Sonnet 在本研究进行时代表了美国生成式 AI 的前沿，在多个方面超越了 GPT-4。选用这 2 个模型旨在比较中美两国在生成式 AI 领域的最新技术进展，并在放射学报告生成的具体应用场景中评估它们的性能差异。

本研究使用 Python 3.8.16 脚本和 OpenAI 1.33.0 包与 ERNIE 4.0 和 Claude 3.5 Sonnet 的应用程序编程接口（application programming interface，API）进行交互。将所有模型的温度参数设置为 1×10^{-10} 以限制随机性[20]。考虑到单次输出的令牌长度限制，对每个案例进行了多次对话迭代，以确保生成完整和连贯的影像学结论。用于此过程的代码已公开发布在 GitHub 平台（https：//github.com/lichao312214129/code_for_generating_impression）。使用正则表达式从 AI 模型的输出中提取影像学结论。

1.3 提示工程

受近期关于如何创建优秀放射学报告建议研究的启发[21]，对初始提示实施了逐步优化提示以提高质量和透明度。这种思维链方法模拟了放射科医师的推理过程，通过结构化分析引导 AI。迭代过程涉及每个案例的多次对话，并在每一步进行人工验证以确保内容的完整性、准确性和连贯性。所有提示采用中文以匹配中国放射科医师的临床环境和患者的目标受众。通过反复优化提示结构和语言（例如开始时可能只是 1 个简单的指令，如“根据报告的描述部分生成报告的结论部分”，发现效果不佳后改为“逐一列出描述中的所有异常，然后结合所有异常生成相应的影像学结论”），使生成式 AI 生成的影像学结论的质量和可靠性得到显著提高，为后续比较分析提供坚实基础。

1.4 性能评估

由 5 名放射科医师（3 名分别具有 17 年、9 年、9 年工作经验的中级医师及 2 名具有 4 年工作经验的初级医师）对纳入研究的 300 份影像学报告（每一份中均包含生成式 AI 与放射科医师的结论）进行独立评估。为确保评估的一致性，对 5 名评估者进行了校准练习：随机选择 5 个案例，评估者分别对其进行独立评估，然后开会讨论结果并形成统一标准。这一过程旨在正式评估前提高评估者间的一致性。

本研究分析了 300 份影像学报告，将其随机分为 2 个子集：子集 1 包含 250 份报告，子集 2 包含 50 份报告。将子集 1 随机分配给 5 名评估者，每名评估者独立评估 50 份报告。将子集 2 分配给所有评估者进行重复评估，即 5 名评估者均对子集 2 的 50 份报告进行独立评估。这一设计旨在测量评估者之间的一致性。在评估过程中，对除影像学报告外的其他临床数据进行盲法处理。为比较生成式 AI 和放射科医师的表现，将子集 1（250 份报告）的评估数据与年资最高的评估者对子集 2（50 份报告）的评估数据整合在一起，创建一个包含 300 份报告的综合数据集用于进一步分析。这种一致性分析方法在以往文献中已有相关报道[22]。

为减少偏倚，每份影像学报告（无论是生成式 AI 的结论还是放射科医师的结论）均被分配由 7 个随机字符串构成的唯一标识符。对于每个案例，ERNIE 4.0、Claude 3.5 Sonnet 和放射科医师的结论的顺序被随机化，以防止任何与顺序相关的评估偏倚。报告以随机顺序呈现的方式在以往的研究中已有报道[20]。

评估采用五点 Likert 量表（1 表示强烈不同意，5 表示强烈同意），重点关注 5 个关键标准：完整性、幻觉、准确性、表达和无修改接受度，每个指标的详细评分标准已公开发布在 GitHub 平台（https：//github.com/lichao312214129/code_for_generating_impression）。这种评估方法参考了既往研究中的实践经验[23]。

1.5 统计学方法

所有统计分析使用 Python 3.8.16、SciPy 1.10.1、Scikit-posthocs 0.8.1 和 Statsmodels 0.13.5 进行[24]。对连续变量（如年龄和报告字数）采用 M（P_{25}，P_{75}）描述，对分类变量（如性别、患者来源、检查方式、增强情况和检查部位）采用 n（%）描述，以全面概括 300 例研究对象的基本特征分布。由于某些标准（如幻觉）的评分分布极端，大多数案例被 5 名评估者一致评为 1，常规的评估者间一致

性测量标准（如 Fleiss’ Kappa 系数）不适用，因此，对于每项评估标准（完整性、幻觉、准确性、表达和无修改接受度）进行评估者间评分一致性的占比计算，一致性被定义为 3 个层级：完全一致（5 位评估者给出相同评分），高度一致（4 位评估者给出相同评分）和基本一致（3 位评估者给出相同评分），比较采用χ^2 检验。

采用 Friedman 检验比较 ERNIE 4.0、Claude 3.5 Sonnet 和放射科医师在 5 个评估标准上的表现。使用 Nemenyi 检验进行事后成对比较（每项评估标准评分采用$\bar{x} \pm s$表示）。采用双侧检验，$P < 0.05$ 为差异有统计学意义。为确保可重复性，所有用于数据分析和可视化的代码同样公开发布在 GitHub 平台（https：//github.com/lichao312214129/code_for_generating_impression）。

2 结 果

2.1 一般资料

本研究分析了 300 例患者的 300 份 CT 和 MRI 报告，300 例患者以中年男性为主，检查部位以上腹部为主，主要来源于住院部和门诊。报告包括 164 份 CT 扫描和 136 份 MRI 扫描，其中 253 份为增强检查。影像学所见部分描述的字数中位数为 320 字。患者一般资料见表 1。

表 1 300 例接受 CT 和 MRI 检查患者的一般资料

Table 1 General information of 300 patients receiving CT and MRI examinations

项 目	数值
性别 /n（%）	
男	202（67.3）
女	98（32.7）
年龄[a]/ 岁	52.0（40.0，62.0）
患者来源 /n（%）	
急诊	28（9.3）
住院	139（46.3）
门诊	133（44.3）
检查方式 /n（%）	
CT	164（54.7）
MR	136（45.3）
增强检查 /n（%）	
对比增强	253（84.3）
无对比增强	47（15.7）
检查部位 /n（%）	
上腹部	238（79.3）
中 + 下腹部	33（11.0）
下腹部	26（8.7）
中腹部	1（0.3）
上 + 中腹部	1（0.3）
上 + 中 + 下腹部	1（0.3）
影像学报告字数	320（255，403）

注：[a]1 例患者的年龄数据缺失。

2.2 5 名评估者之间的一致性

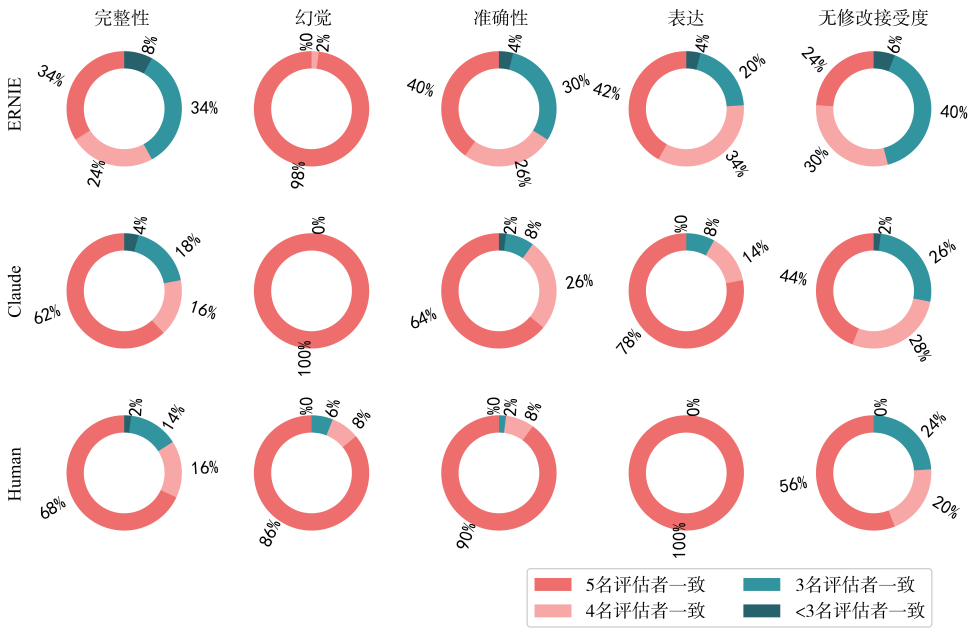
所有评估标准显示评估者间一致性均较高，见图 2。对于 ERNIE 4.0 生成的结论，至少 3/5 的评估者评分一致的占比情况：完整性（92.0%），幻觉（100%），准确性（96.0%），表达（96.0%），无修改接受度（94.0%）。对于 Claude 3.5 Sonnet 生成的结论，至少 3/5 的评估者评分一致的占比情况：完整性（96.0%），幻觉（100%），准确性（98.0%），表达（100%），无修改接受度（98.0%）。对于放射科医师的结论，至少 3/5 的评估者评分一致的占比情况：完整性（98.0%），幻觉（100%），准确性（100%），表达（100%），无修改接受度（100%）。ERNIE 4.0、Claude 3.5 Sonnet 和放射科医师在完整性（$\chi^2 = 12.59$，$P < 0.01$）、幻觉（$\chi^2 = 12.59$，$P < 0.01$）、准确性（$\chi^2 = 12.24$，$P < 0.01$）、表达（$\chi^2 = 24.32$，$P < 0.01$）和无修改接受度（$\chi^2 = 21.58$，$P < 0.01$）5 个指标比较差异均有统计学意义。

2.3 性能比较

Nemenyi 检验显示，在完整性和无修改接受度方面，ERNIE 4.0 的得分低于 Claude 3.5 Sonnet 和放射科医师（P 均 = 0.001），后两者之间比较差异无统计学意义（P 均 > 0.05）。在准确性方面，放射科医师优于 ERNIE 4.0 和 Claude 3.5 Sonnet（P 均 = 0.001）。在幻觉和表达方面，3 组表现相似，比较差异均无统计学意义（P 均 > 0.05），见图 3。

总体而言，Claude 3.5 Sonnet 在多个方面的表现与放射科医师相当，而 ERNIE 4.0 在某些领域仍有改进空间。Friedman 检验显示幻觉在 3 组间的差异具有统计学意义，但进一步采用 Nemenyi 检验进行两两比较，并未发现任意 2 组间的差异具有统计学意义，这可能是由于大多数幻觉得分为 1、分布高度偏斜所致。生成式 AI 与放射科医师在影像学报告生成任务中的表现见表 2。

本研究发现人类医师生成的影像学报告也存在“幻觉”现象，即影像学结论包含了影像学所见中未描述的内容。这主要源于放射科报告的审核流程：资深医师在审核初级医师的报告时，可能发现初级医师遗漏的重要病变，但出于对工作效率的考虑，资深医师往往在影像学结论部分直接



注：该图展示了评估者对生成式 AI 和放射科医师的结论达成一致评分的百分比，凸显了评估过程的可靠性和一致性。

图 2　5 名评估者间的一致性

Figure 2　Consistency among 5 evaluators

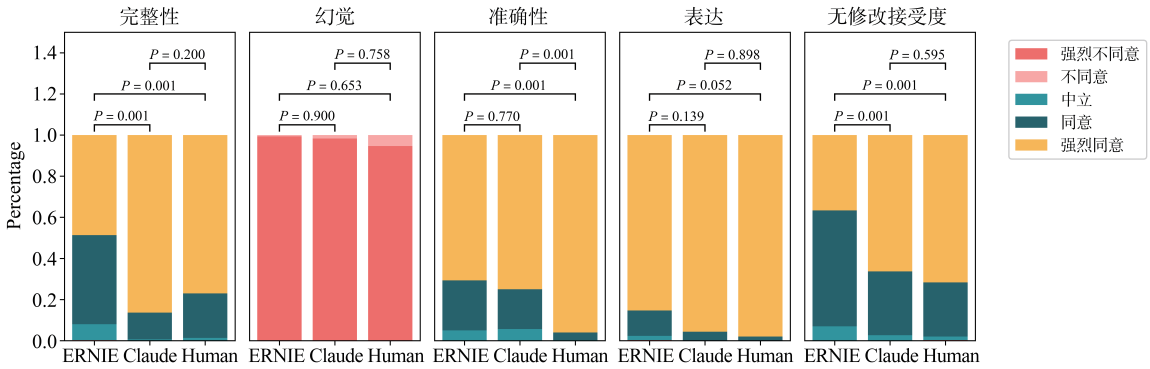


图 3　ERNIE 4.0、Claude 3.5 Sonnet 和放射科医师结论的 5 项评估标准比较

Figure 3　Comparison of five evaluation criteria for ERNIE 4.0, Claude 3.5 Sonnet and radiologist conclusions

表 2　AI 模型与放射科医师在医学影像报告生成任务中的表现比较　　单位：分

Table 2　Comparison of AI models and radiologists' performance in medical image report generation tasks

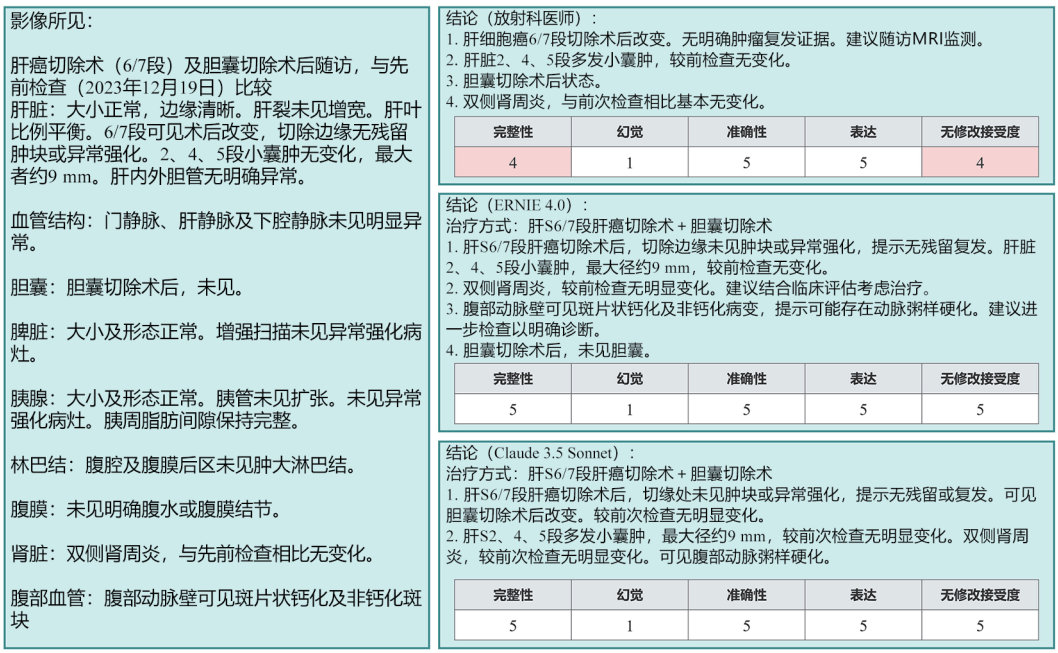
指　标	ERNIE 4.0	Claude 3.5 Sonnet	放射科医师
完整性	4.40 ± 0.64	4.86 ± 0.37	4.76 ± 0.46
幻觉性	1.01 ± 0.08	1.02 ± 0.13	1.05 ± 0.23
准确性	4.66 ± 0.57	4.69 ± 0.57	4.96 ± 0.22
表达力	4.82 ± 0.48	4.96 ± 0.20	4.98 ± 0.14
修改度	4.30 ± 0.59	4.64 ± 0.53	4.69 ± 0.54

添加这些内容，而未更新影像学所见部分，导致两者出现不一致的情况。

ERNIE 4.0 和 Claude 3.5 Sonnet 以及放射科医师的结论案例见图 4、5。

3　讨　论

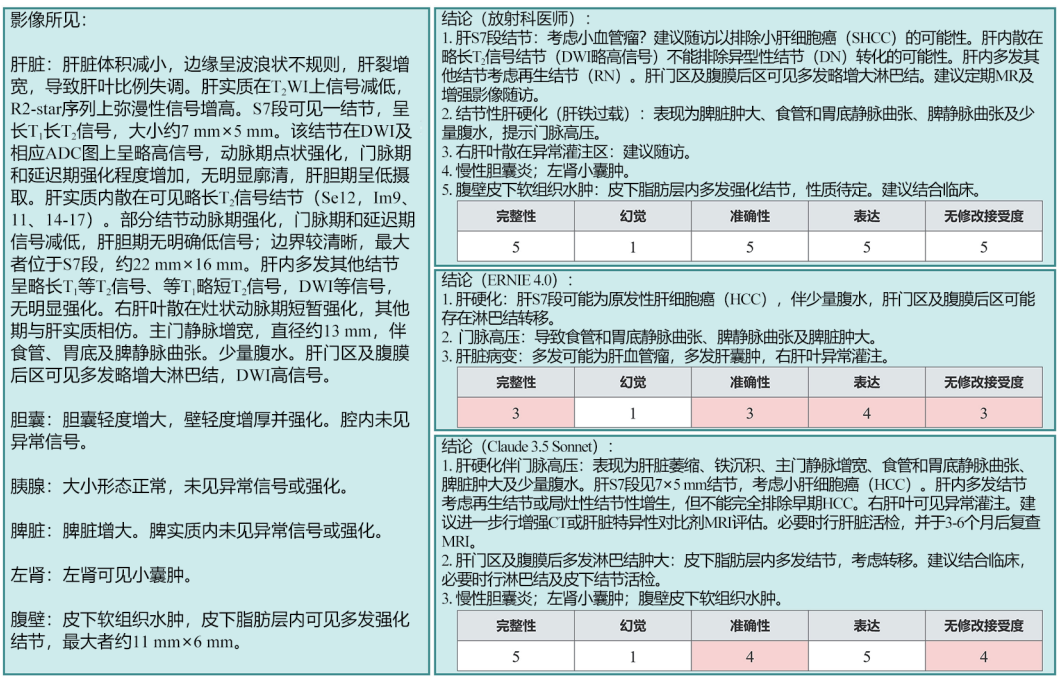
本研究比较了 2 种先进的生成式 AI 在生成腹部 CT 和 MRI 影像学报告方面的表现，并与放射科医师进行比较。结果表明，Claude 3.5 Sonnet 在多个方面达到了与放射科医师相当的水平，而 ERNIE 4.0 则显示出有改进空间，尤其是在完整性和无修改接受度方面，ERNIE 4.0 的表现劣于



注：左图为影像报告的影像学所见部分。右图包括 3 个子图，第 1 个子图显示放射科医师的结论，其中遗漏了主动脉粥样硬化的诊断；第 2 个子图显示 ERNIE 4.0 生成的结论，包括所有相关发现，如主动脉粥样硬化；第 3 个子图显示 Claude 3.5 Sonnet 生成的结论，同样包括所有相关发现。

图 4　ERNIE 4.0、Claude 3.5 Sonnet 和放射科医师的影像学报告案例一

Figure 4　ERNIE 4.0, Claude 3.5 Sonnet and radiologist imaging report Case 1



注：左图显示影像学报告的影像学所见部分。右图包括 3 个子图，第 1 个子图显示放射科医师的结论；第 2 个子图显示 ERNIE 4.0 生成的结论，其中遗漏了关键诊断，如慢性胆囊炎、左肾囊肿、腹壁皮下软组织水肿和皮下脂肪层多发强化结节，此外还将肝血管瘤误诊为肝细胞癌，将肝硬化结节误诊为多发血管瘤；第 3 个子图显示 Claude 3.5 Sonnet 生成的结论，虽然也将肝 S7 段的血管瘤误诊为肝细胞癌，但准确包含了其他相关发现。

图 5　ERNIE 4.0、Claude 3.5 Sonnet 和放射科医师的影像学报告案例二

Figure 5　ERNIE 4.0, Claude 3.5 Sonnet and radiologist imaging report Case 2

Claude 3.5 Sonnet。三者在表达和幻觉方面表现相似。然而在准确性方面，放射科医师优于 Claude 3.5 Sonnet 和 ERNIE 4.0。上述结果表明了 Claude 3.5 Sonnet 在生成影像学报告方面具有较强能力，同时也表明当前由生成式 AI 生成影像学报告应在放射科医师的监督下完成。

Claude 3.5 Sonnet 的表现在某些方面优于 ERNIE 4.0，这可能归因于训练数据、模型架构或每个模型的提示工程效果的差异。这凸显了将 AI 应用于放射学时模型选择和优化的重要性[20,23]。

本研究的结果不仅验证了既往研究的部分发现，还在多个方面拓展了现有的认知范围。与 Sun 等[13]的研究结果不同，本研究显示，在多个评估维度上，先进的 AI 模型能够生成与放射科医师质量相当的结论，这一差异可能与 AI 模型在性能上的提升以及本研究采用了更复杂的提示工程技术有关。与 Nakaura 等[16]的研究相比，本研究进行了更全面的评估，Nakaura 等仅纳入具有典型影像学特征的 28 份报告，而本研究的影像学报告涵盖了更广泛的疾病类型，不限于典型特征。这种方法更有助于评估 AI 在处理各种复杂和不同表现形式病例时的性能。

本研究所用方法更接近 Ziegelmayer 等[18]的建议，强调专业放射科医师评估在评价 AI 生成内容方面的重要性。与 Gundogdu 等[17]主要依赖自动化指标的研究相比，本研究进行了更具临床相关性的评估。此外，本研究在样本规模和影像学模态多样性方面具有显著优势，涵盖了大量 CT 和 MRI 检查结果。这种研究设计不仅增强了统计分析的可靠性，还有助于全面评估 AI 模型在处理不同解剖复杂度和成像模态时的性能。通过这种多维度的分析方法，本研究深入探讨了 2 种生成式 AI 在应对腹部影像学报告中各类挑战时的能力，从而提供了更为细致和全面的见解。

本研究存在一定局限性。首先，作为单中心研究，本研究结果的泛化性可能有限，需要多中心研究来验证这些结果在不同临床环境中的适用性。其次，本研究专注于腹部影像学报告，可能无法反映 AI 在其他检查部位的表现，未来的研究应扩展至更全面的人体检查部位。第三，虽然本研究的评估标准较全面，但开发更细化的指标可能更有助于对 AI 模型在特定临床环境中的表现进行深入分析。最后，未来应开展纵向研究，以评估 AI 模型在临床实践中的整合对放射科医师诊断的准确性、工作效率以及患者临床结局的长期影响。

本研究结果表明，生成式 AI 在放射学工作流程中展现出显著潜力，有望成为一种有价值的辅助工具。本研究也揭示了一些需要进一步优化和改进的部分。随着 AI 技术的持续发展，将其有效整合到放射学临床实践中可能会显著提升工作效率，并有望改善患者的管理质量。未来在放射学领域融入 AI 模型时，应着重关注放射科医师与 AI 模型协作模式的构建，而非单纯追求技术替代。这种协作应充分发挥 AI 的计算能力和放射科医师的临床经验，以此优化诊断流程，提高整体诊断的准确性和效率。此外，还需要进行更多的前瞻性研究，以评估 AI 技术辅助诊断在实际临床环境中的长期效果和影响。

参 考 文 献

[1] MASKELL G. Why does demand for medical imaging keep rising[J]. BMJ, 2022, 379: o2614.DOI: 10.1136/bmj.o2614.

[2] LAI A Y T. The growing problem of radiologist shortage: Hong Kong's perspective[J]. Korean J Radiol, 2023, 24(10): 931-932. DOI: 10.3348/kjr.2023.0838.

[3] RAWSON J V, SMETHERMAN D, RUBIN E. Short-term strategies for augmenting the national radiologist workforce[J]. AJR Am J Roentgenol, 2024, 222(6): e2430920.DOI: 10.2214/ajr.24.30920.

[4] VOSSHENRICH J, BRANTNER P, CYRIAC J, et al. Quantifying radiology resident fatigue: analysis of preliminary reports[J]. Radiology, 2021, 298(3): 632-639.DOI: 10.1148/radiol.2021203486.

[5] ALEXANDER R, WAITE S, BRUNO M A, et al. Mandating limits on workload, duty, and speed in radiology[J]. Radiology, 2022, 304(2): 274-282. DOI: 10.1148/radiol.212631.

[6] 许溪，康宁，罗敏婷，等. 人工智能在儿童耳鼻咽喉头颈外科中应用的系统综述[J]. 新医学，2024，55(7)：497-505. DOI：10.3969/j.issn.0253-9802.2024.07.002.
XU X, KANG N, LUO M T, et al. Application of artificial intelligence in pediatric otolaryngology-head and neck surgery: a systematic review[J]. J New Med, 2024, 55(7): 497-505. DOI: 10.3969/j.issn.0253-9802.2024.07.002.

[7] 诸露冰，汪建华. 医学影像人工智能在胰腺癌精准诊疗中的研究进展[J]. 新医学，2024，55(3)：153-158. DOI：10.3969/j.issn.0253-9802.2024.03.001.
ZHU L B, WANG J H. Research progress on medical imaging-based artificial intelligence in precision diagnosis and treatment of pancreatic cancer[J]. J New Med, 2024, 55(3): 153-158. DOI: 10.3969/j.issn.0253-9802.2024.03.001.

[8] BHAYANA R. Chatbots and large language models in radiology:

a practical primer for clinical and research applications[J]. Radiology, 2024, 310(1): e232756. DOI: 10.1148/radiol.232756.

[9] MOOR M, BANERJEE O, ABAD Z S H, et al. Foundation models for generalist medical artificial intelligence[J]. Nature, 2023, 616(7956): 259-265. DOI: 10.1038/s41586-023-05881-4.

[10] HASANI A M, SINGH S, ZAHERGIVAR A, et al. Evaluating the performance of Generative Pre-trained Transformer-4(GPT-4) in standardizing radiology reports[J]. Eur Radiol, 2024, 34(6): 3566-3574. DOI: 10.1007/s00330-023-10384-x.

[11] MESE I, TASLICAY C A, SIVRIOGLU A K. Improving radiology workflow using ChatGPT and artificial intelligence[J]. Clin Imaging, 2023, 103: 109993. DOI: 10.1016/j.clinimag.2023.109993.

[12] 秦江涛, 王继荣, 肖一浩, 等. 人工智能在医学领域的应用综述[J]. 中国医学物理学杂志, 2022, 39(12): 1574-1578. DOI: 10.3969/j.issn.1005-202X.2022.12.019.
QIN J T, WANG J R, XIAO Y H, et al. Artificial intelligence in medical application: a review[J]. Chin J Med Phys, 2022, 39(12): 1574-1578. DOI: 10.3969/j.issn.1005-202X.2022.12.019.

[13] SUN Z, ONG H, KENNEDY P, et al. Evaluating GPT4 on impressions generation in radiology reports[J]. Radiology, 2023, 307(5): e231259.DOI: 10.1148/radiol.231259.

[14] RAY P P.The need to re-evaluate the role of GPT-4 in generating radiology reports[J]. Radiology, 2023, 308(2): e231696. DOI: 10.1148/radiol.231696.

[15] MA C, WU Z, WANG J, et al. An iterative optimizing framework for radiology report summarization with ChatGPT[J]. IEEE Trans Artif Intell, 2024, 5(8): 4163-4175. DOI: 10.1109/TAI.2024.3364586.

[16] NAKAURA T, YOSHIDA N, KOBAYASHI N, et al. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports[J]. Jpn J Radiol, 2024, 42(2): 190-200. DOI: 10.1007/s11604-023-01487-y.

[17] GUNDOGDU B, PAMUKSUZ U, CHUNG J H, et al. Customized impression prediction from radiology reports using BERT and LSTMs[J]. IEEE Trans Artif Intell, 2021, 4(4): 744-753. DOI: 10.1109/TAI.2021.3086435.

[18] ZIEGELMAYER S, MARKA A W, LENHART N, et al. Evaluation of GPT-4's chest X-ray impression generation: a reader study on performance and perception[J]. J Med Internet Res, 2023, 25: e50865. DOI: 10.2196/50865.

[19] KIM W. Seeing the unseen: advancing generative AI research in radiology[J]. Radiology, 2024, 311(2): e240935. DOI: 10.1148/radiol.240935.

[20] BHAYANA R, NANDA B, DEHKHARGHANIAN T, et al. Large language models for automated synoptic reports and resectability categorization in pancreatic cancer[J]. Radiology, 2024, 311(3): e233117. DOI: 10.1148/radiol.233117.

[21] HARTUNG M P, BICKLE I C, GAILLARD F, et al. How to create a great radiology report[J]. Radiographics, 2020, 40(6): 1658-1670. DOI: 10.1148/rg.2020200020.

[22] 钟丽茹, 罗娜, 唐文杰. 双能量CT电子云密度和有效原子序数在甲状腺良恶性结节鉴别诊断中的价值[J]. 新医学, 2024, 55(9): 716-721.DOI: 10.3969/j.issn.0253-9802.2024.09.006.
ZHONG L R, LUO N, TANG W J. The value of dual-energy CT electron cloud density and effective atomic number in differential diagnosis of benign and malignant thyroid nodules[J]. J New Med, 2024, 55(9): 716-721.DOI: 10.3969/j.issn.0253-9802.2024.09.006.

[23] FINK M A, BISCHOFF A, FINK C A, et al. Potential of ChatGPT and GPT-4 for data mining of free-text CT reports on lung cancer[J]. Radiology, 2023, 308(3): e231362. DOI: 10.1148/radiol.231362.

[24] POLLARD T J, JOHNSON A E W, RAFFA J D, et al. Tableone: an open source Python package for producing summary statistics for research papers[J]. JAMIA Open, 2018, 1(1): 26-31. DOI: 10.1093/jamiaopen/ooy012.

(责任编辑: 洪悦民)