

Translation for AJR article “*Decoupling Visual Parsing and Diagnostic Reasoning for Vision–Language Models (GPT-4o and GPT-5): Analysis Using Thoracic Imaging Quiz Cases*”

by lead author **Dae Hee Han** from Department of Radiology, Seoul St. Mary's Hospital

Translation by **Xin Bai** from Department of Radiology, Peking Union Medical College Hospital

标题：解耦视觉解析与诊断推理在视觉语言模型（GPT-4o 与 GPT-5）中的应用：基于胸部影像测试病例的分析

第一作者： Dae Hee Han 首尔圣玛丽医院 放射科

翻译及朗读： 白鑫 中国医学科学院北京协和医院 放射科

背景： 视觉语言模型具备识别放射影像表现（即视觉解析）并将发现转化为诊断（即诊断推理）的潜力。然而，现有模型表现尚不足以支持临床整合。

目的： 评估视觉解析和诊断推理分别对基于 GPT 的视觉语言模型在胸部影像中生成正确诊断的性能贡献。

方法： 本回顾性研究纳入了来自韩国胸部影像学会测试平台的 128 例公开胸部影像病例（获取于 2025 年 6 月 15 日）。评估了两种视觉语言模型（GPT-4o 和 GPT-5）分别在输入患者元数据和影像、以及输入患者元数据和放射科医生生成的影像描述时的表现。模型为每个病例提供五个按排名排序的鉴别诊断；当输入元数据和影像时，模型首先提供影像发现的总结。计算了模型的五个鉴别诊断中包含正确诊断的病例比例（即前五位准确率）。从平台提取了使用元数

据和影像解读病例的测试参与者的表现。对模型提供的影像总结质量采用 4 分制评分（4 分为最佳）。采用逻辑回归分析评估模型影像总结评分与诊断性能之间的关联。评估了模型的首要诊断与测试参与者的前十位鉴别诊断列表之间的一致性。

结果：GPT-4o 和 GPT-5 在输入元数据和影像时的前五位准确率分别为 15.9% 和 24.7%，在输入元数据和影像描述时分别为 40.1% 和 59.1%；测试参与者汇总的前五位准确率为 45.8%。两种模型的影像总结评分中位数均为 2；这些评分显示出与进入前五位匹配的显著独立关联（GPT-4o，OR=5.95；GPT-5，OR=2.77； $P<.001$ ）。GPT-4o 和 GPT-5 在输入元数据和影像时，其首要诊断与测试参与者鉴别诊断列表的一致性分别为 31.6% 和 39.3%，在输入元数据和影像描述时分别为 78.8% 和 79.4%。

结论：两种视觉语言模型在视觉识别胸部影像表现方面能力有限，但在获得放射科医生生成的描述后，其生成准确诊断的表现更为良好。

临床意义：研究结果强调了放射科医生在胸部影像解读中的专业价值，并指出限制视觉语言模型性能的主要瓶颈在于视觉图像解析能力，而非诊断推理能力。