DOI:10.3969/j.issn.1005-202X.2025.11.019

医学人工智能

评估通用与领域特定大视觉模型在食管癌CT的T分期中的应用：一项关于零样本性能及提示词工程影响的研究

朱大兵[1],高伟[2],林杨皓[1],赖武浩[3],梁志超[3],曾宪一[2],邓希楷[2],安军[3,4]
1.中山大学附属第三医院粤东医院放射科，广东 梅州 514700；2.联通数智医疗科技有限公司，广东 广州 511457；3.中山大学附属第三医院粤东医院心胸外科，广东 梅州 514700；4.中山大学附属第三医院心胸外科，广东 广州 510630

【摘要】背景：食管癌精准T分期对治疗至关重要，但CT评估局限明显。大视觉模型（LVMs）提供了新可能，但其未经微调的“零样本”临床诊断能力尚待验证。**方法：**本研究回顾性分析98例食管癌及50例正常对照的胸部CT影像，在放射科专家共识的金标准下，通过不同复杂度的提示词，评估GPT-5、Gemini及MedGemma的零样本T分期性能。**结果：**GPT-5表现出最高的准确性与稳定性。模型间存在显著偏倚：Gemini倾向过度分期，MedGemma则倾向分期不足。所有模型对早期肿瘤判断均不佳，但结构化提示词能提升对中晚期病灶的诊断性能。**结论：**LVMs具备零样本T分期的潜力，但性能高度依赖模型选择与提示词设计。通用模型GPT-5的零样本泛化能力更优。当前模型性能，尤其在早期诊断上，远未达到临床应用标准。未来方向在于高质量数据微调与标准化提示框架开发。

【关键词】食管癌；大视觉模型；CT分期；提示词工程；领域专用模型；零样本学习；诊断准确性
【中图分类号】R318；R735.1 **【文献标志码】**A **【文章编号】**1005-202X(2025)11-1532-09

Evaluating generic and domain-specific large visual models for T staging of esophageal cancer using CT: a study of zero-shot performance and the impact of prompt engineering

ZHU Dabing[1], GAO Wei[2], LIN Yanghao[1], LAI Wuhao[3], LIANG Zhichao[3], ZENG Xianyi[2], DENG Xikai[2], AN Jun[3,4]
1. Department of Radiology, Yuedong Hospital, the Third Affiliated Hospital of Sun Yat-sen University, Meizhou 514700, China; 2. China Unicom Digital Intelligent Medical Technology Co., Ltd., Guangzhou 511457, China; 3. Department of Cardiothoracic Surgery, Yuedong Hospital, the Third Affiliated Hospital of Sun Yat-sen University, Meizhou 514700, China; 4. Department of Cardiothoracic Surgery, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China

Abstract: Background Accurate T-staging is critical for esophageal cancer therapy, but CT-based assessment has significant limitations. Large vision models (LVMs) hold promise, yet their zero-shot clinical diagnostic capability without fine-tuning remains unvalidated. **Methods** A retrospective analysis was conducted on the chest CT images from 98 esophageal cancer patients and 50 normal controls. Using radiologist-consensus as the gold standard, the zero-shot T-staging performance of 3 LVMs (GPT-5, Gemini, and MedGemma) was evaluated with prompts of varying complexity. **Results** GPT-5 exhibited the highest accuracy and stability. Significant biases were observed among models: Gemini tended to over-stage, while MedGemma showed a tendency to under-stage. All models faced challenges in identifying early-stage tumors, but structured prompts improved diagnostic performance for mid-to-late stage lesions. **Conclusion** LVMs have potential for zero-shot T-staging, but their performance highly depends on model choice and prompt design. The generic model GPT-5 show superior zero-shot generalization. However, current model performance is not yet clinically viable, especially for early diagnosis. Future work should focus on fine-tuning with high-quality clinical data and developing standardized prompt frameworks.

Keywords: esophageal cancer; large vision model; computed tomography staging; prompt engineering; domain-specific model; zero-shot learning; diagnostic accuracy

【收稿日期】2025-09-13
【基金项目】教育部产学合作协同育人项目（22087043260712）；梅州市医学科研项目（2024C0302005）
【作者简介】朱大兵，副主任医师，研究方向：胸部、甲状腺影像学诊断，E-mail: 13549143903@163.com；高伟，硕士，研究方向：医学大数据与人工智能，E-mail: gaow@chinaunicom.cn
【通信作者】安军，博士，副主任医师，研究方向：胸部肿瘤、肿瘤免疫，E-mail: anjun@mail.sysu.edu.cn

0 引言与背景

0.1 食管癌的临床挑战：流行病学与精准分期的迫切性

食管癌是全球范围内最常见的消化道恶性肿瘤之一，其发病率和死亡率均居高不下，对人类健康构成严重威胁[1]。根据全球癌症统计数据，食管癌是第8大常见癌症，位列癌症相关死亡原因的第6位[2]。该疾病主要包括两种组织学亚型：食管鳞状细胞癌（ESCC）和食管腺癌（EADC）。这两种亚型在地理分布、流行病学特征及主要风险因素上表现出显著差异[1]。ESCC在东亚至中亚的“食管癌带”以及东非部分地区高发，其风险因素与贫困、吸烟、饮酒、热饮及营养缺乏等密切相关[2]。相比之下，EADC在西方国家更为常见，其发病率在过去几十年中呈上升趋势，主要与胃食管反流病（GERD）、Barrett食管、肥胖和吸烟等因素有关[1]。

在食管癌的临床管理中，一个无法回避的核心挑战是其预后与诊断时的疾病分期紧密相连。基于美国癌症联合委员会（AJCC）的TNM分期系统，对肿瘤的局部侵犯范围（T分期）、区域淋巴结转移（N分期）和远处转移（M分期）进行精确评估，是制定个体化治疗策略、预测患者生存结局的决定性因素[3]。临床数据显示，早期局限性食管癌（0~I期）患者的5年生存率可达45%~50%，而一旦疾病进展至区域转移（II~III期），生存率则降至20%~30%；对于发生远处转移（IV期）的患者，5年生存率更是低于5%[1]。这种巨大的预后差异凸显了精准分期的极端重要性。准确的分期不仅能够筛选出适合接受根治性手术或内镜下切除的早期患者，还能为局部晚期患者确定最佳的新辅助放化疗方案，同时避免为已发生广泛转移的患者施加不必要的局部治疗创伤。因此，任何能够提升TNM分期准确性的技术，都可能对改善食管癌患者的临床结局产生深远影响。

0.2 食管癌分期中的CT影像：既定标准及其内在局限

在当前临床实践中，增强多排螺旋计算机断层扫描（CT）是食管癌术前分期的基石性影像学检查手段[3]。CT以其快速、无创、可重复性强以及能够提供全身解剖信息的优势，在评估肿瘤大小、位置、与邻近纵隔结构的关系、区域淋巴结状态以及肝肺等远端器官转移方面发挥着不可或缺的作用，为多学科团队制定治疗决策提供了关键依据[4]。然而，尽管CT技术不断进步，其在食管癌精准分期，特别是T分期评估中的固有局限性也日益凸显，这些局限性构成了当前诊断精度的主要瓶颈。

首先，CT在区分早期T分期方面能力有限。由于其空间分辨率的限制，CT无法清晰分辨食管壁的精细解剖层次结构（如黏膜层、黏膜下层、肌层和外膜）。这一缺陷导致CT在鉴别肿瘤局限于黏膜或黏膜下层（T1期）与侵犯肌层（T2期）时表现出根本性的不可靠性[5]。这一区分对于治疗决策至关重要，因为T1a期病变可能适用于创伤更小的内镜下切除，而T2期及以上病变则通常需要更为激进的外科手术或新辅助治疗。一项包含868例CT分期患者的大型回顾性研究量化了这一挑战，结果显示CT对病理T1期和T2期肿瘤的诊断准确率分别仅为53%和63%[6]。

其次，对于局部晚期病变，CT在评估肿瘤对邻近结构侵犯（T3/T4期）的准确性也存在不确定性。虽然CT能够较好地显示肿瘤穿透外膜侵犯纵隔脂肪组织（T3期），但在判断微小脂肪浸润时仍缺乏可靠性[5]。更具挑战性的是T4期的判断，即肿瘤是否侵犯了主动脉、气管、支气管或椎体等不可切除的结构。临床上，肿瘤与邻近器官之间脂肪间隙的消失是提示侵犯的重要征象，但这一征象的特异性并不高。在恶病质或体型偏瘦的患者中，正常的解剖间隙可能本身就很窄；而在接受过放疗的患者中，放射性炎症和纤维化也可能导致脂肪间隙模糊，这些情况都可能导致T4期的假阳性诊断，从而错误地剥夺患者接受手术治疗的机会[5]。尽管有研究提出了一些量化指标，如肿瘤与主动脉接触弧度超过90°，但这些标准在临床实践中被证实并不可靠[5]。上述研究同样指出，尽管CT对T3和T4期的诊断准确率分别提升至89%和93%，但仍存在一定比例的误判[6]。这些局限性共同构成了一个关键的临床难题：我们依赖一种存在明确“天花板”的工具来做出影响患者生死的重大决策。这为探索能够超越传统影像解读模式、提供更精准分期信息的新技术创造了强烈的临床需求。

0.3 肿瘤影像人工智能（AI）的演进：从狭义智能到通用基础模型

AI在放射学和肿瘤成像领域的应用，经历了一场深刻的技术演进，正从解决单一、特定任务的“狭义AI”向具备多任务、多模态理解能力的“通用AI”范式转变。早期，AI在医学影像中的成功应用主要集中于深度学习，特别是卷积神经网络（CNNs）的部署[7]。这些模型在特定任务上表现出色，例如，在食管癌领域，研究人员开发了专门用于CT或内镜图像中病灶自动检测、分割以及基于影像组学特征预测淋巴结转移或治疗反应的算法[8]。这些“专家模型”通常需要大量经过精细标注的、特定于任务的数据集进行监督式学习，其功能高度特化，泛化到新任务

或不同类型数据时能力有限[9]。然而，近年来AI领域最引人注目的突破是基础模型（Foundation Models）的崛起，尤其是在自然语言处理领域以大型语言模型（LLMs）和在视觉领域以大型视觉模型（LVMs）为代表的技术[10]。这些模型与传统狭义AI的核心区别在于其训练范式和能力范围。基础模型通过在海量的、多样化的、通常是未标注的数据上进行自监督学习，构建了对世界（包括文本、图像、声音等）的广泛、通用的理解[11]。这种预训练使其获得了强大的“涌现能力”，即在没有针对性训练的情况下，通过少量的样本（few-shot）甚至零样本（zero-shot）的方式，解决各种复杂的下游任务[12]。这一范式转变正在重塑医学AI的未来，催生了“通用医学人工智能”（Generalist Medical AI, GMAI）的概念[13]。GMAI模型旨在灵活地处理和整合来自不同来源的医疗数据，如影像、电子病历、基因组学和病理报告，并能生成富有表现力的输出，如自由文本的诊断解释或图像标注，从而展现出高级的医学推理能力[13]。在医学影像领域，这意味着模型可能不再仅仅是识别像素模式的工具，而是能够结合图像特征与内置的或通过提示词提供的医学知识，进行类似放射科医生的认知过程-观察、描述、分析和诊断[11]。这一技术飞跃为解决前述CT分期的局限性提供了一条全新的路径，即利用LVMs强大的视觉理解和推理能力，直接从CT图像中解读出更精确的T分期信息，而无需依赖传统算法所需的大规模像素级标注。

0.4 研究的基本原理、问题与目标

综合以上背景，本研究的出发点在于一个关键的、尚未被充分探索的交叉领域。一方面，我们面临着食管癌CT分期准确性不足的长期临床困境，这一困境直接影响患者治疗决策和预后。另一方面，我们正处在一个由通用基础模型驱动的AI技术革命浪潮中，这些模型理论上具备解决复杂视觉推理任务的潜力。然而，一个核心的未知问题是：这些在通用数据（如互联网图片和文本）上训练出的“通才”模型，其所学的知识和能力是否能够有效地迁移到高度专业化、充满细微差别且对错误容忍度极低的临床医学影像诊断任务中？特别是在零样本条件下，即模型在从未见过特定医院、特定设备、特定人群的数据集上，其表现如何？这种从通用领域到专科领域的知识迁移并非理所当然。医学影像的解读不仅需要识别形态学特征，更需要将其置于复杂的解剖学、病理生理学和临床背景中进行综合判断。因此，本研究旨在填补这一认知空白，通过一个具体而关键的临床问题——食管癌CT的T分期来检验LVMs的真实世界应用潜力。

本研究旨在回答以下3个核心问题：（1）零样本诊断能力评估：通用大视觉模型（GPT-5、Gemini等）和领域专用模型（MedGemma）在未经本地数据微调的情况下，能否对一个全新的、本地化的食管癌CT数据集进行具有临床参考价值的T分期诊断？其准确性能否与已知的传统CT诊断水平相比较？（2）提示词工程的影响：用户与AI模型的交互方式-即提示词的设计对诊断性能有多大影响？从简单指令到模拟专家工作流程的复杂结构化提示词，是否能够系统性地提升模型的诊断准确性和推理的可靠性？（3）模型间性能差异分析：不同的LVMs（通用型 *vs* 专用型，不同技术架构）在处理同一任务时，是否会展现出独特的性能特征或系统性偏倚（如倾向于过度分期或分期不足）？

基于上述问题，本研究的目标是：设计并实施一项初步的、探索性的对比评估研究，系统地考察3款代表性的LVMs在3种不同复杂度的提示词驱动下，对本地食管癌CT影像进行零样本T分期诊断的性能。通过这项研究，期望为LVMs在肿瘤学领域的未来发展方向、临床整合策略以及人机交互模式的优化提供基础性证据和深刻见解。

1 方 法

1.1 研究设计与患者队列

研究方案获得了机构审查委员会的批准，并本研究为单中心、回顾性研究，纳入2023年~2025年8月间148例患者的胸部增强CT影像，其中，食管癌组98例（II期7例，III期48例，IV期43例），均有手术病理结果作为T分期依据。另纳入50例正常CT影像作为对照组。

1.2 金标准的确立

为了给模型的性能评估建立一个可靠、客观的参照标准（即“金标准”），本研究采用严格的多人独立评估与共识机制。所有入组的148例CT影像（包括98例食管癌和50例正常对照）均被匿名化处理，并分发给3位具有5年以上胸部肿瘤影像诊断经验的主治级别放射科医生。评估过程遵循“背对背”（back-to-back）的盲法原则，在独立评估完成后，收集3位医生的诊断结果，对于模型识别肿瘤存在与否的评估，以3位医生一致的影像学诊断为准。

1.3 评估的LVMs

为了全面考察不同类型基础模型在医学影像零样本任务中的表现，本研究策略性地选择了3款具有代表性的大视觉模型。这3款模型的选择旨在覆盖从通用型到领域专用型、从现有顶尖模型到未来前沿模型的不同维度，从而能够对模型架构和预训练数据的影响进行初步探讨。

（1）GPT-5（OpenAI）。本研究中使用的GPT-5是一个基于其前代模型（如GPT-4V）性能轨迹和公开技术路线的前瞻性、假设性模型。选择它代表了当前通用多模态模型的最高技术水平和未来发展方向[14-15]。

（2）Gemini（Google）。Gemini强调原生多模态能力的深度融合，能够无缝地处理和推理文本、图像、音频和视频等多种信息[16]。用以考察不同技术路线下的通用模型在同一项专业医学任务上的性能异同，从而增强研究结果的普适性[17]。

（3）MedGemma（Google）。MedGemma是Google基于其Gemini架构，专门为医疗领域开发的专用基础模型[18]。与通用模型不同，MedGemma在Gemini的通用能力基础上，利用海量的、经过匿名化处理的医疗数据（包括医学文献、电子病历和医学影像）进行了进一步的微调和优化[18]。

通过对比这3款模型，本研究能够深入分析通用知识泛化与领域知识特化在医学影像AI应用中的相对优势和潜在挑战。

1.4 提示词工程策略

提示词工程，即设计和优化输入给AI模型的指令，是决定模型输出质量和相关性的关键环节。大量研究表明提示词的结构、内容和清晰度能显著影响大型模型的推理路径和最终性能[19]。为了系统性地探究这种影响，本研究设计了3个层次递进的提示词，分别代表简单、标准和复杂的交互模式。

（1）简单提示词（Simple Prompt）。内容："识别图中异常结构，如果存在肿瘤进行影像学TNM分期诊断"。

（2）标准提示词（Standard Prompt）。内容："请你作为一名专业的AI放射科医生，分析我上传的食管癌患者增强CT影像。请详细描述肿瘤的位置、形态、尺寸和强化特征，并依据影像进行严谨的TNM分期：重点评估肿瘤的侵犯深度（T分期），特别是与主动脉、气管等周围结构的脂肪间隙关系；识别并报告所有可疑的区域转移淋巴结（N分期）；同时系统性地检查肝、肺等器官有无远处转移（M分期）。最后，虽然CT无法确定病理分级，请根据肿瘤的侵袭性特征（如坏死、边界等）对其恶性程度进行推断，并将你的完整分析以一份结构化的报告输出。在报告第一句进行明确的TNM分期诊断"。

（3）复杂提示词（Complex Prompt）。内容由角色、背景信息、输出格式、影像分析摘要、详细发现、分析依据部分构成的详细提示词。

通过对比这3种提示词下的模型表现，本研究能够量化评估提示词复杂度对LVMs在专业医学影像任务中诊断准确性、报告质量和推理可靠性的影响。

1.5 数据分析与性能指标

将148张CT影像分别输入3个LVMs，并依次使用3种提示词进行查询，构成9个独立实验组。主要评估指标包括：（1）肿瘤识别率：模型在98例食管癌影像中成功识别出肿瘤的比例。（2）T分期诊断准确率：模型给出的T分期与病理金标准完全一致的比例。所有数据采用描述性统计进行分析。

2 结　果

2.1 基础模型能力（病灶识别与分期可行性）

在评估模型进行复杂T分期诊断的准确性之前，首要步骤是确定它们是否具备两项基础能力：一是在CT影像中成功识别出肿瘤病灶的存在，二是在识别后能够执行分期这一指定任务。研究结果显示，在病灶识别层面，大多数模型-提示词组合表现出极高的鲁棒性。具体而言，GPT-5和MedGemma在所有3种提示词下，均100%成功地在98例食管癌影像中识别出肿瘤结构。然而，Gemini模型在与简单提示词组合时表现出明显的不足，未能识别出其中的8例肿瘤病例，识别失败率达到8.2%。当Gemini与标准或复杂提示词组合时，其识别能力得到改善，能够识别所有98例肿瘤。这一发现初步揭示了Gemini模型对提示词的依赖性相对更强，简单的指令不足以稳定激活其病灶检测能力。

在分期可行性层面，即模型识别肿瘤后是否能提供T分期诊断，观察到更显著的差异。GPT-5模型表现出最佳的稳定性，在所有98例识别出的肿瘤病例中，无论使用何种提示词，均100%成功地给出T分期判断。相比之下，Gemini模型再次显示出不稳定性，即使在识别出肿瘤后，仍有相当比例的病例无法完成分期任务。具体来说，使用简单提示词时，有7例无法判断T分期；使用标准提示词时有4例无法判断；使用复杂提示词时仍有6例无法判断。MedGemma模型在简单提示词下也存在2例无法判断T分期的情况，但在标准和复杂提示词下则能对所有识别出的肿瘤进行分期。这些"分期失败"的案例表明，部分模型在从视觉感知（识别异常）到临床推理（进行分期）的认知链条中存在断点。

2.2 总体T分期诊断准确性与模型性能特征

本研究的核心目标是评估LVMs在零样本条件下的T分期诊断准确性。综合分析所有9个实验组的结果，观察到清晰的模型性能梯度和独特的诊断偏倚模式。图1详细汇总了各模型-提示词组合在98例食管癌病例中的总体表现。

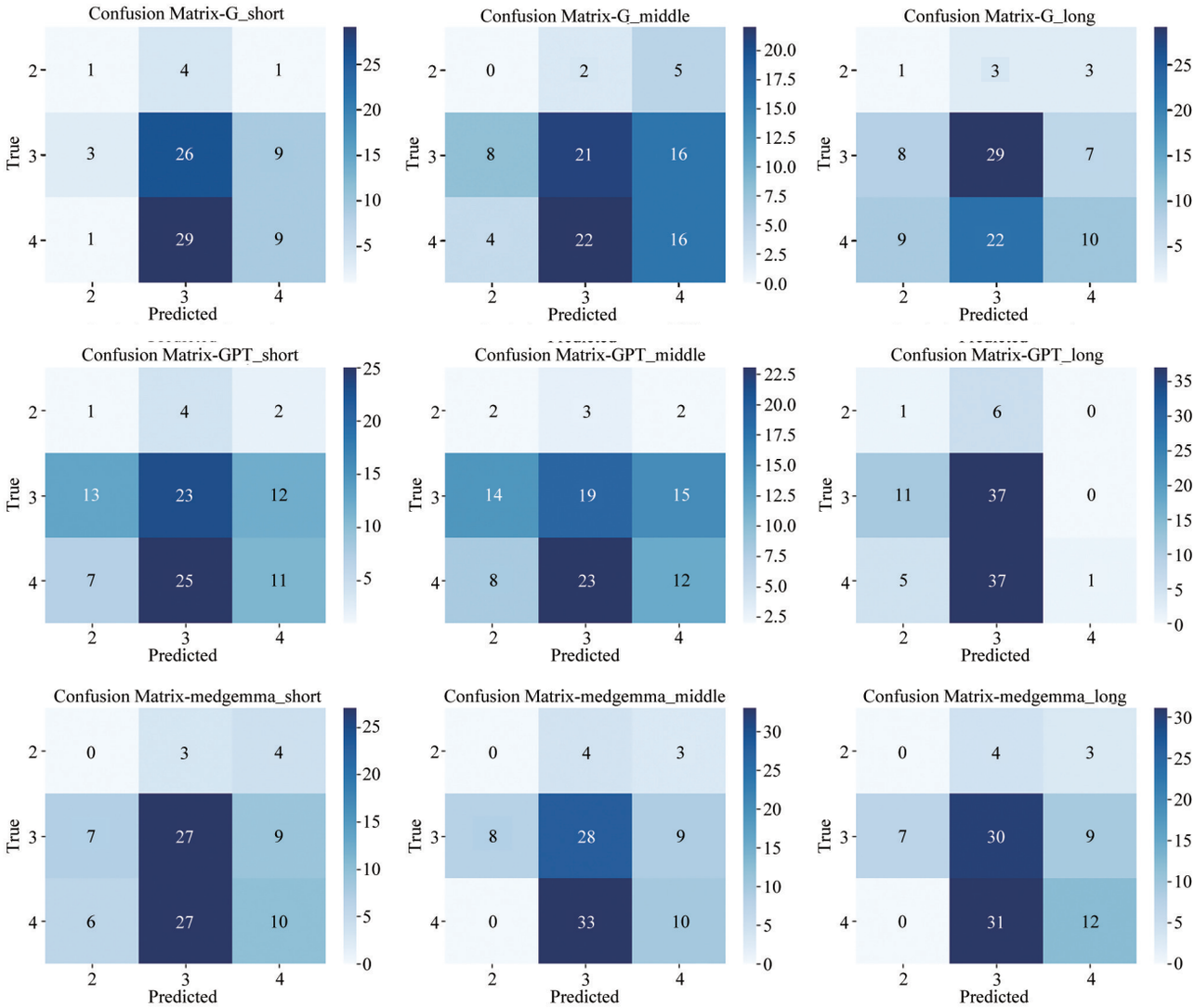

图1 大视觉模型在不同提示词下的总体T分期性能特征

Figure 1 Overall T-staging performance characteristics of large vision models under different prompts

2.3 不同模型T分期的性能分析

为了更深入地理解模型的行为模式，合并不同提示词情况下的数据，单纯分析不同模型的T分期诊断准确性。图2展示这一分层分析的结果。从判断的整体来看，Gemini对影像的分期评估的T分期水平较低，可能与其有多个影像图片未识别出肿瘤结构相关。ChatGPT对T2级的图像判断能力明显较差，未识别出T2的CT影像。而MedGemma作为医学方向的垂类大模型，其判断T分期的结果整体偏高，同时可以识别出一定的T2期结果，合理性较其他两个大模型更强。

进一步结合人工判断的结果作为标准，评估准确性。结果发现相对而言，GPT-5在所有模型中展现出的整体性能较弱，更为有趣的是，标准模式的提示词对GPT-5反而起到了负向影响，而全面化的提示词对GPT-5的T分期准确率有明显提示。Gemini模型的整体准确性居中，而提示词的完整程度与其判断的准确率有着明显的相关性，这表明其强大的通用视觉理解和推理能力能够较好地迁移到专业的医学影像判读任务中。MedGemma作为领域专用模型，其总体准确性在本研究的零样本测试中最高，这与其经过专项的大模型优化有着必然的联系。相较于其他两个通用模型的规模，我们只在本地部署了一个7B规模的MedGemma模型就获得了相较于其他大模型T分期更高的准确率。

总体来看，模型间的性能层级清晰地呈现为：MedGemma>Gemini>GPT-5。这一结果提示，在当前的零样本场景下，顶尖通用大模型的泛化能力不足以超越参数规模相对较小或训练数据不够泛化的领域专用模型，但能力接近。

2.4 提示词复杂度对诊断准确性的影响

本研究的一个核心探索点是提示词对LVMs诊

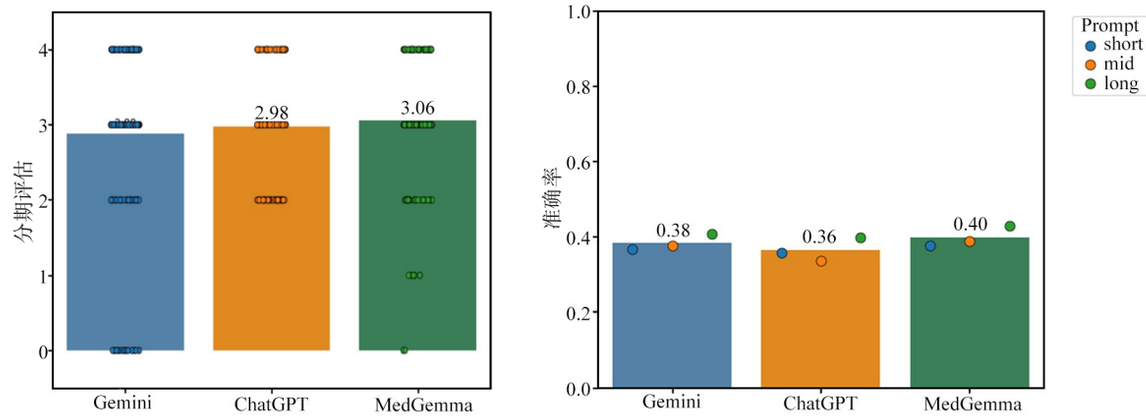

图2 不同模型对CT图像的T分期判断趋势以及分期诊断准确性

Figure 2 Trends in T-staging judgment and the accuracy of staging diagnosis of different models for CT images

断性能的调节作用。通过对比简单、标准、复杂3种提示词下的结果，发现了一个一致且明确的趋势：增加提示词的结构性和信息量能够普遍提升模型的诊断准确性。短提示词的整体准确率37%，而长提示词整体准确率达到41%，这个结果支持了上面的判断。但是提示词在不同模型中可能的效果会有差别。对于MedGemma和Gemini这两个模型而言，当提示词升级为标准提示词和复杂提示词时，两个模型的性能均得到显著改善。这些更详细的提示词通过设定角色、分解任务和明确评估标准，为模型提供了一个清晰的“思维框架”，有效地引导其分析过程。对于GPT-5而言，在中等长度提示词情况下的准确率下降，让人感到意外，这可能说明GPT-5模型可能对提示词的细节更为敏感，需要更详尽、更结构化的指令才能充分发挥其潜力。

总而言之，研究结果有力地证明了提示词在人与AI协作进行医学影像诊断中的关键价值。精心设计的、模拟临床专家工作流程的提示词，是解锁和规范化LVMs诊断能力的重要工具。

3 讨 论

3.1 主要发现的综合分析：领域特化与人机交互共同定义AI性能

本研究的核心贡献在于：在一个具体的、具有挑战性的临床场景中，系统性地评估了前沿LVMs在未经任何本地数据微调（即零样本）的情况下，进行食管癌CT的T分期诊断的能力。主要发现颠覆了“通用模型性能至上”的普遍认知，揭示领域专用模型在高度专业化任务中的显著优势，并强调人机交互（即提示词工程）的复杂性和非线性效应。

首先，研究证实了领域专用模型MedGemma的卓越性能。尽管其模型参数规模（7B）远小于顶尖的通用模型，但在T分期这一专业任务上却取得了最高的准确率。这一结果有力地表明在医学影像解读这类需要精细特征识别和深厚领域知识的任务中，经过医疗数据预训练所带来的知识内化，其价值可能超过了通用模型由更大参数和更泛数据所带来的泛化能力[20]。通用模型如GPT-5和Gemini虽然在基础的病灶识别上表现尚可，但在关键的T分期判断上，尤其是对于传统CT诊断的难点（如T2期），则暴露出明显短板，如GPT-5完全未能识别任何T2期影像，这暗示其通用视觉知识无法有效转化为对食管壁微小侵犯层次的理解。

其次，本研究揭示了模型从“感知”到“推理”的认知链条中可能存在的断点。Gemini和MedGemma在简单提示词下，即便识别出肿瘤，也偶有无法完成分期任务的情况。这表明模型的视觉感知能力与其临床推理能力并非总是同步激活，特别是在缺乏明确指令引导时，模型可能无法稳定地将其视觉发现转化为符合临床逻辑的诊断结论。

最后，提示词工程的关键作用得到了证实，但其影响并非简单的线性关系。总体而言，更结构化、更详细的提示词能提升诊断准确性，这与鼓励模型进行“思维链”（Chain-of-Thought）推理能改善其性能的发现一致[4]。然而，GPT-5在标准提示词下准确率不升反降的反常现象，为我们敲响了警钟：提示词工程并非“万金油”，其效果高度依赖于特定模型的内部架构和“思维模式”。这表明优化人机交互远比想象中复杂，需要针对不同模型进行精细化的设计与验证[21]。

3.2 领域专用模型的胜利：知识深度战胜规模广度

本研究最引人注目的发现是，领域专用模型MedGemma在零样本T分期任务中全面胜出，其性能层级清晰地呈现为MedGemma>Gemini>GPT-5。这

一结果对当前基础模型发展中“越大越好”的理念提出了挑战，并凸显了领域特化预训练（domain-specific pre-training）在专业应用中的不可替代性[22]。

通用大模型（如GPT-5和Gemini）通过在海量互联网数据上训练，学习了广泛的世界知识和视觉模式[10]。然而，医学影像与自然图像之间存在巨大的领域鸿沟[23]。CT图像的灰度、纹理、解剖结构和病理特征具有高度的特异性，这些细微的差别对于通用模型而言可能是噪音，但对于诊断却至关重要。MedGemma的成功，正因为它通过在海量的、去识别化的医疗数据（包括影像、报告和临床记录）上进行预训练或微调，已经将这些医学领域的“先验知识”编码到了模型参数中[18]。因此，它在解读食管癌CT时，并非从零开始，而是基于一个已经“理解”医学影像语言的基础。其能够识别出部分T2期病变（尽管整体分期偏高），而GPT-5则完全失败，这正是领域知识深度的体现。

相比之下，通用模型的表现则印证了通用知识在专科领域的局限性。尽管它们可能在某些方面（如文本问答）超越人类医生[17]，但在需要精细视觉解读的任务上，其性能并不稳定[14]。GPT-5的整体表现不佳，以及Gemini的性能与提示词高度相关，均表明它们在面对一个陌生的专业任务时，更像一个依赖外部指令的“通才”，而非一个具备内在专业直觉的“专家”。这一发现与一些研究相呼应，即直接将在ImageNet等自然图像数据集上预训练的模型迁移到医学影像任务，效果往往不尽如人意[20]。因此，对于高风险、高专业的医疗应用，发展和应用领域专用基础模型，应是比单纯追求通用模型规模更有效、更安全的路径。

3.3 提示词工程的非线性效应：引导、依赖与“GPT-5之谜”

本研究证实提示词工程是解锁和规范化LVMs诊断能力的关键工具，但其作用机制远比“指令越清晰越好”更为复杂和微妙。

对于Gemini和MedGemma，我们观察到一种相对线性的、符合预期的关系：随着提示词从简单到标准再到复杂，其结构性、角色扮演和任务分解的元素逐渐增多，模型的诊断准确率也随之稳步提升。这与大量研究的结论一致，即采用“思维链”（Chain-of-Thought）或模拟专家工作流程的结构化提示，能够有效引导模型进行更深入、更可靠的推理[4]。这表明这两款模型能够很好地利用外部提供的“脚手架”来组织其内部的分析过程，其性能表现出对高质量引导的积极响应。然而，GPT-5的表现则揭示了提示词工程的非线性乃至反常的效应。标准（中等复杂度）提示词反而导致其准确率下降，这一“反常现象”是本研究一个极具启发性的发现。对此，我们提出几种可能的解释。第一，指令冲突或误解：标准提示词中的某些表述可能与GPT-5内置的、从通用数据中学到的推理模式相冲突，导致其“无所适从”。第二，过度约束：对于一个极其强大的通用模型，过于具体但又不够完美的指令，可能反而限制了其自由探索和发现正确推理路径的能力，不如简单的指令给予其更大的“发挥空间”。第三，模型敏感性：这可能反映了GPT-5这类顶尖模型对提示词的细微变化具有极高的敏感性，微小的措辞差异都可能将其引导至完全不同的“思维”轨道。

无论原因为何，“GPT-5之谜”有力地说明，我们不能将提示词工程简单地视为一种通用的优化技术，它更像是一种与特定模型“性格”和“知识结构”进行交互的艺术。未来的研究必须从“通用提示词技巧”转向“针对特定模型的个性化交互策略”，深入理解不同模型架构如何响应不同的指令模式，这对于构建稳定、可靠的人机协同诊断系统至关重要[24]。

3.4 本地化AI的曙光：“小而美”模型的战略价值

本研究中一个在本地部署的、仅有7B参数的MedGemma模型战胜了云端的、参数规模可能大出数倍的通用巨头，这一结果为医疗机构的AI战略提供了极具价值的启示：发展和部署“小而美”的本地化AI模型，可能是一条比依赖外部通用大模型更具成本效益、更安全也更有效的路径。

这一发现的核心在于数据和模型的协同价值。MedGemma的成功，本质上是高质量的领域数据（用于其预训练）与高效模型架构结合的胜利。这为各医疗机构指明了方向：其自身积累的高质量、结构化的临床数据，是无可比拟的战略资产。通过利用这些本地数据，对一个像MedGemma这样的领域专用基础模型进行进一步的微调，有望实现性能的再次飞跃[25]。这种“基础模型+本地数据微调”的模式具有多重优势：（1）性能更优：微调能够使模型适应本院特有的患者人群、疾病谱、影像设备参数和诊疗习惯，从而获得比任何通用模型都更高的诊断精度[26]。（2）成本更低：训练和推理一个7B规模的模型，其计算资源需求远低于动辄千亿参数的通用模型，使得院内部署成为可能，降低了对昂贵云服务的依赖[27]。（3）数据更安全：将模型部署在本地，意味着敏感的患者数据无需离开医院的防火墙，从根本上解决了数据隐私和安全的核心关切。（4）应用更可控：医疗机构可以对本地化模型进行持续的监控、验证和迭代，确保其性能的稳定性和可靠性，避免了因外部模型版本更新而导致的性能波动。

因此，本研究的结果不仅是对MedGemma模型的一次验证，更是对一种未来医疗AI发展范式的有力支持，即从追求“大而全”的通用智能，转向构建“小而精”、与本地临床深度融合的专用智能。这要求医疗机构不仅要成为AI技术的使用者，更要成为高质量数据的治理者和本地化AI模型的构建者[28]。

3.5 临床意义、未来方向与研究局限

3.5.1 临床意义　本研究的结果清晰地表明，尽管领域专用模型展现出巨大潜力，但当前所有受测的大视觉模型，均未达到可独立用于临床诊断的水平。其在T2期食管癌诊断上的普遍低准确率，以及不同模型中观察到的系统性偏倚（如MedGemma的分期偏高），都构成了不可接受的潜在安全风险。模型的输出可能产生“幻觉”或基于错误的推理[14]，若直接用于指导治疗，可能导致对早期患者的过度治疗或对晚期患者的治疗不足。然而，MedGemma的优异表现预示着经过进一步的本地化微调和严格验证后，这类模型极有潜力成为放射科医生的得力助手。

3.5.2 未来方向　基于本研究的发现和局限，未来的研究应沿着以下几个方向深入：(1)本地数据微调的实证研究。最为迫切的下一步是利用本研究的本地数据集，对表现最佳的MedGemma模型进行监督式微调（SFT）乃至直接偏好优化（DPO）等更先进的训练[25]。通过量化对比微调前后的性能，可以直接验证“本地化是通往临床级性能的关键”这一核心假设。(2)扩大样本与多中心验证。本研究的单中心设计和有限的样本量（尤其是II期病例）限制了结论的普适性。未来的研究必须纳入更大规模、来自不同医疗中心、使用不同品牌CT设备的数据，以测试和提升模型的泛化能力和鲁棒性[29]。(3)多模态数据融合。临床诊断本身就是一个多模态过程。未来的AI模型不应局限于2D CT切片，而应能处理完整的3D容积数据，并整合来自PET-CT的功能代谢信息、电子病历中的临床文本信息（如内镜报告、病史），甚至基因组学数据，以构建一个更全面、全景式的患者模型，从而做出更精准的决策[30]。(4)临床工作流的无缝整合。如何将这些强大的AI工具高效且安全地整合到繁忙的放射科日常工作流中，是一个巨大的技术和实践挑战[31]。未来的研究需要关注人机交互界面设计、结果呈现方式以及如何处理AI生成的意外发现等实际问题，确保技术能够真正落地并提升临床效率，而非增加负担[32]。

3.5.3 研究局限性　本研究存在几个明确的局限性。首先，其回顾性和单中心设计可能引入选择偏倚，且结果未必能外推至其他医疗机构。其次，样本量相对较小，特别是II期病例仅有7例，这使得对早期肿瘤诊断性能的结论仅为初步观察。第三，我们采用了2D轴位图像进行评估，这简化了输入，但牺牲了放射科医生在阅片时利用冠状位和矢状位进行三维空间判断所能获得的丰富信息。

4 结 论

本研究对通用与领域专用大视觉模型在食管癌CT的T分期中的零样本诊断能力进行了初步但系统的探索。研究结果表明这些前沿的AI模型，特别是顶尖的通用模型，确实具备了在无需特定训练的情况下解读复杂医学影像并进行临床推理的潜在能力。然而，这种能力尚处于初级阶段，其性能表现出对模型架构和提示词设计的强烈依赖性，且在诊断准确性、稳定性和处理早期病变的能力方面，与临床独立应用的要求尚有巨大差距。

本研究最重要的启示在于，通向临床级医学影像AI的道路，可能并非寄望于一个“开箱即用”的通用超级智能。更现实、更可靠的路径是一种协同进化的模式，即以强大的、经过医学领域优化的基础模型为起点，通过精心设计的、模拟临床专家思维的提示词工程进行有效引导，并最终利用高质量、结构化的本地临床数据进行深度微调。这种结合了强大基础架构、精妙人机交互和关键本地数据的协同策略，将是未来开发出真正安全、可靠并能为肿瘤患者带来实际价值的AI诊断工具的核心所在。本项工作为未来方向提供了基础性的证据支持，并强调了在AI技术浪潮中临床专业知识和高质量医疗数据不可替代的核心价值。

【参考文献】

[1] Yang H, Wang F, Hallemeier CL, et al. Oesophageal cancer[J]. Lancet, 2024, 404(10466): 1991-2005.

[2] Zhang YW. Epidemiology of esophageal cancer [J]. World J Gastroenterol, 2013, 19(34): 5598-5606.

[3] Hong SJ, Kim TJ, Nam KB, et al. New TNM staging system for esophageal cancer: what chest radiologists need to know [J]. Radiographics, 2014, 34(6): 1722-1740.

[4] Niekel MC, Bipat S, Stoker J. Diagnostic imaging of colorectal liver metastases with CT, MR imaging, FDG PET, and/or FDG PET/CT: a meta-analysis of prospective studies including patients who have not previously undergone treatment[J]. Radiology, 2010, 257(3): 674-684.

[5] Quint LE, Bogot NR. Staging esophageal cancer[J]. Cancer Imaging, 2008, 8 Spec No A(Spec Iss A): S33-S42.

[6] Zhao WJ, Huang J, Chen YK, et al. The clinical value of CT and MRI in preoperative TNM staging of esophageal cancer[J]. Front Med (Lausanne), 2025, 12: 1637764.

[7] Zheng SY, Cui XN, Ye ZX. Integrating artificial intelligence into radiological cancer imaging: from diagnosis and treatment response to prognosis[J]. Cancer Biol Med, 2025, 22(1): 6-13.

[8] Fu J, Huang XY, Fang MJ, et al. Artificial intelligence in medical imaging empowers precision neoadjuvant immunochemotherapy in esophageal squamous cell carcinoma[J]. J Immunother Cancer, 2025, 13(9): e012468.

[9] Li SW, Zhang LH, Cai Y, et al. Deep learning assists detection of esophageal cancer and precursor lesions in a prospective, randomized

controlled study[J]. Sci Transl Med, 2024, 16(743): eadk5395.
[10] Fang MJ, Wang ZP, Pan ST, et al. Large models in medical imaging: advances and prospects[J]. Chin Med J (Engl), 2025, 138(14): 1647-1664.
[11] Wang R, Chen ZS. Large-scale foundation models and generative AI for BigData neuroscience[J]. Neurosci Res, 2025, 215: 3-14.
[12] Hou WP, Liu Q, Ma HF, et al. Assessing large multimodal models for one-shot learning and interpretability in biomedical image classification[J]. Adv Intell Syst.
[13] Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence[J]. Nature, 2023, 616(7956): 259-265.
[14] Jin Q, Chen FY, Zhou YL, et al. Hidden flaws behind expert-level accuracy of multimodal GPT-4 vision in medicine[J]. NPJ Digit Med, 2024, 7(1): 190.
[15] Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis[J]. Eur Radiol, 2025, 35 (4): 1959-1965.
[16] Hirosawa T, Harada Y, Tokumasu K, et al. Comparative study to evaluate the accuracy of differential diagnosis lists generated by Gemini advanced, Gemini, and bard for a case report series analysis: cross-sectional study[J]. JMIR Med Inform, 2024, 12: e63010.
[17] Huang KA, Choudhary HK, Hardin WM, et al. Comparative analysis of ChatGPT-4o and Gemini advanced performance on diagnostic radiology in-training exams[J]. Cureus, 2025, 17(3): e80874.
[18] Zhou SC, Yu S. High-throughput biomedical relation extraction for semi-structured web articles empowered by large language models[J]. BMC Med Inform Decis Mak, 2025, 25(1): 351.
[19] Liu JL, Liu F, Wang CY, et al. Prompt engineering in clinical practice: tutorial for clinicians[J]. J Med Internet Res, 2025, 27: e72644.
[20] Almattar W, Anwar S, Al-Azani S, et al. Multi scale self supervised learning for deep knowledge transfer in diabetic retinopathy grading [J]. Sci Rep, 2025, 15(1): 33742.
[21] Wang L, Chen X, Deng XW, et al. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs[J]. NPJ Digit Med, 2024, 7(1): 41.
[22] Wang ZF, Wang HY, Danek B, et al. A perspective for adapting generalist AI to specialized medical AI applications and their challenges[J]. NPJ Digit Med, 2025, 8(1): 429.
[23] Vendrow E, Schonfeld E. Understanding transfer learning for chest radiograph clinical report generation with modified transformer architectures[J]. Heliyon, 2023, 9(7): e17968.
[24] Wada A, Akashi T, Shih G, et al. Optimizing GPT-4 turbo diagnostic accuracy in neuroradiology through prompt engineering and confidence thresholds[J]. Diagnostics (Basel), 2024, 14(14): 1541.
[25] Savage T, P Ma S, Boukil A, et al. Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization: comparative evaluation[J]. J Med Internet Res, 2025, 27: e76048.
[26] Yang QM, Chen JX, Sun Y, et al. Fine-tuning medical language models for enhanced long-contextual understanding and domain expertise[J]. Quant Imaging Med Surg, 2025, 15(6): 5450-5462.
[27] Pingua B, Sahoo A, Kandpal M, et al. Medical LLMs: fine-tuning *vs*. retrieval-augmented generation[J]. Bioengineering (Basel), 2025, 12 (7): 687.
[28] de Almeida JG, Alberich LC, Tsakou G, et al. Foundation models for radiology-the position of the AI for health imaging (AI4HI) network [J]. Insights Imaging, 2025, 16(1): 168.
[29] Zhang WY, Chang YJ, Shi RH. Artificial intelligence enhances the management of esophageal squamous cell carcinoma in the precision oncology era[J]. World J Gastroenterol, 2024, 30(39): 4267-4280.
[30] Haq IU, Mhamed M, Al-Harbi M, et al. Advancements in medical radiology through multimodal machine learning: a comprehensive overview[J]. Bioengineering (Basel), 2025, 12(5): 477.
[31] Kotter E, Ranschaert E. Challenges and solutions for introducing artificial intelligence (AI) in daily clinical workflow[J]. Eur Radiol, 2021, 31(1): 5-7.
[32] Korfiatis P, Kline TL, Meyer HM, et al. Implementing artificial intelligence algorithms in the radiology workflow: challenges and considerations[J]. Mayo Clin Proc Digit Health, 2025, 3(1): 100188.

（编辑：黄开颜）