DOI:10.11798/j.issn.1007-1520.202524552

·论著·

颈部单中心型 Castleman 病临床辅助诊疗中 ChatGPT o1 和 Claude 3.5 Sonnet 的应用比较研究

潘鑫[1],郜飞[2],陈亭亭[3],朱一鸣[2],程小凌[1],梁江盟[1],岳天宇[4],张政[5],雷齐鸣[1],卫旭东[1,2,3]

(1.兰州大学第一临床医学院,甘肃 兰州 730000;2.甘肃省人民医院 耳鼻咽喉头颈外科,甘肃 兰州 730000;3.甘肃中医药大学,甘肃 兰州 730000;4.吉林大学,吉林 长春 130000;5.西安交通大学,陕西 西安 710000)

摘 要:目的 研究并对比分析 ChatGPT o1 与 Claude 3.5 Sonnet 在解答颈部单中心 Castleman 病相关常见问题时的差异。**方法** 围绕颈部单中心 Castleman 病设计 36 个常见问题,收集并输入到 ChatGPT o1 与 Claude 3.5 Sonnet 搜索引擎中,由耳鼻咽喉头颈外科学教授分别对 ChatGPT o1 和 Claude 3.5 Sonnet 生成的回答进行独立评估,评估内容涵盖回答内容的可读性、准确性、质量、易理解程度以及实际可操作性。**结果** 在可读性方面,Claude 3.5 Sonnet 在所有类别中生成的回答字数更简短(189.36±69.09 vs. 381.56±153.28,$P<0.05$),具有更低的阅读分数(1.68±5.64 vs. 11.20±11.16,$P<0.05$),年级分数更高(54.93±35.81 vs. 16.70±2.03,$P<0.05$)。在患者教育材料评估工具(PEMAT-P)评分衡量的可理解性和可操作性方面,Claude 3.5 Sonnet 表现出更高的总体可理解性(0.38±0.17 vs. 0.06±0.05,$P<0.05$)和可操作性(0.25±0.22 vs. 0.08±0.09,$P=0.015$)。然而,ChatGPT o1 的总体准确度分数更高(4.88±0.28 vs. 4.58±0.37,$P=0.002\ 2$),并在修改后的基于证据的患者教育信息质量评估工具(EQIP)标准下获得了更好的质量分数(7.47±1.28 vs. 5.75±1.20,$P<0.05$)。**结论** Claude 3.5 Sonnet 在简洁性、可理解性和可操作性方面占优势,而 ChatGPT o1 在准确性、整体质量和可读性上更胜一筹。

关 键 词:颈部;单中心型 Castleman 病;ChatGPT;Claude

中图分类号:R739.91

A comparative study on the application of ChatGPT o1 and Claude 3.5 Sonnet in the clinical adjuvant diagnosis and treatment of unicentric Castleman disease in the neck

PAN Xin[1], GAO Fei[2], CHEN Tingting[3], ZHU Yiming[2], CHENG Xiaoling[1], LIANG Jiangmeng[1], YUE Tianyu[4], ZHANG Zheng[5], LEI Qiming[1], WEI Xudong[1,2,3]

(1. *the First School of Clinical Medicine, Lanzhou University, Lanzhou* 730000, *China*; 2. *Department of Otolaryngology Head and Neck Surgery, Gansu Provincial Pepole's Hospital, Lanzhou* 730000, *China*; 3. *Gansu University of Chinese Medicine, Lanzhou* 730000, *China*; 4. *Jilin University, Changchun* 130000, *China*; 5. *Xi'an Jiaotong University, Xi'an* 710000, *China*)

Abstract: Objective To study and compare the differences between ChatGPT o1 and Claude 3.5 Sonnet in answering common questions related to unicentric Castleman disease in the neck. **Methods** A total of 36 common questions related to unicentric Castleman disease of the neck were designed and input into the ChatGPT o1 and Claude 3.5 Sonnet search engines. The answers generated by ChatGPT o1 and Claude 3.5 Sonnet were independently evaluated by a professor of otolaryngology-head and neck surgery. The evaluation contents included the readability, accuracy, quality, understandability, and practical operability. **Results** Regarding readability, Claude 3.5 Sonnet generated significantly more concise responses across all categories (189.36±69.09 vs. 381.56±153.28, $P<0.05$), with lower reading scores (1.68±5.64 vs. 11.20±11.16, $P<0.05$) and higher grade-level scores (54.93±35.81 vs. 16.70±2.03, $P<$

基金项目:国家自然科学基金地区科学基金项目(82260475);甘肃省卫生健康行业科技创新重大行业项目(GSWSZD2024-02);甘肃省科技计划重点研发项目(25YFWA028)。

第一作者简介:潘鑫,男,在读硕士研究生,住院医师。

通信作者简介:卫旭东,男,博士,主任医师。

0.05). In terms of understandability and actionability measured by the patient education materials assessment tool-patient (PEMAT-P) score, Claude 3.5 Sonnet demonstrated higher overall understandability (0.38 ± 0.17 vs. 0.06 ± 0.05, $P < 0.05$) and actionability (0.25 ± 0.22 vs. 0.08 ± 0.09, $P = 0.015$). Nevertheless, ChatGPT o1 had a higher overall accuracy score (4.88 ± 0.28 vs. 4.58 ± 0.37, $P = 0.002$ 2) and attained a superior quality rating under the revised evidence-based quality indicator for patient education (EQIP) criteria (7.47 ± 1.28 vs. 5.75 ± 1.20, $P < 0.05$). **Conclusion** Claude 3.5 Sonnet is superior in conciseness, understandability, and actionability, whereas ChatGPT o1 is stronger in terms of accuracy, overall quality and readability.

Keywords: Neck; Unicentric Castleman disease; ChatGPT; Claude

人工智能(artificial intelligence,AI)是一个致力于通过计算机技术模拟和扩展人类智能的领域。它不仅可以独立运作,还能作为人类能力的有效补充,在医疗、金融等众多专业领域发挥重要作用。AI 在医学领域的应用可追溯至 20 世纪 50 年代,经过数十年的发展已取得显著进展[1]。2024 年 12 月,OpenAI 公司发布了 ChatGPT 的 o1 版本;Anthropic(美国加利福尼亚州)开发了大型语言 AI 模型 Claude 3.5 Sonnet,此类 AI 聊天机器人能够理解用户输入并生成智能化的回应,展现出强大的自然语言处理能力。

Castleman 病又称巨大淋巴结病或血管滤泡性淋巴结增生症,是一种罕见的疾病,该病于 1956 年首次由 Castleman 等[2]命名。Castleman 病的发病机制尚未完全阐明,目前普遍认为人类疱疹病毒 8 型(human herpesvirus 8,HHV-8)感染、人类免疫缺陷病毒(human immunodeficiency virus, HIV)感染和白细胞介素-6(interleukin-6, IL-6)的异常表达是与其发病机制相关的主要因素[3-4],对于头颈部单中心型 Castleman 病,暂时没有发现较为独特的因素。Castleman 病在临床上根据病变累及部位分为单中心型 Castleman 病、多中心型 Castleman 病[5]。

随着 ChatGPT 和 Claude 等聊天机器人的兴起,以及人们对医疗信息即时获取需求的增加,AI 正逐渐成为广泛使用的在线医疗信息资源。因此,评估 ChatGPT 和 Claude 在回答常见医疗问题时的答案质量显得尤为重要。然而,关于 ChatGPT 和 Claude 在头颈部单中心 Castleman 病患者问题上的应用研究仍然较少。在本研究中,我们评估并比较了 ChatGPT 和 Claude 生成的关于头颈部单中心 Castleman 病常见问题的回答,从可读性、质量、内容准确程度、易理解程度以及实际可操作性方面对其进行了分析。

1　方法

1.1　研究设计

由于所有获取的信息均为公开可用信息,本研究获得了机构审查委员会的豁免。研究者在各自的官方网站上创建了 ChatGPT o1 与 Claude 3.5 Sonnet 的账号。基于指南《单中心型 Castleman 病国际循证共识诊断和治疗指南》[5]和卡斯尔曼病协作网站(https://cdcn.org),设计 36 个问题,进一步细分为子类别,即基础知识、诊断、治疗,这种分层基于每个问题的相应内容,详情如下:①基础知识(13 个问题):包含疾病的基本信息,包括病因与风险、病理特征与分级、遗传与分子机制、疾病位置与并发症、临床特征与分类、研究与应对措施、相关病症和临床试验。②诊断(8 个问题):涉及疾病诊断中采用的方法,包括评估单中心 Castleman 病的最佳活检方法、临床和实验室评估、全面检查项目和鉴别诊断。③治疗(15 个问题):涉及疾病的治疗和管理策略。对于每个问题,都收集了 ChatGPT o1 与 Claude 3.5 Sonnet 响应,在每次提问之前和之后(包括重复提问之间),都删除了会话历史,以确保每次对话是独立的,不受前一次对话的影响,其中可读性评估、可理解性和可操作性评估、准确性评估从基础知识部分、诊断相关部分、治疗相关部分以及总体情况对两种 AI 模型的文本进行了全面的对比评估。

1.2　可读性评估

为了评估文本的可读性,本文使用 Flesch 阅读容易度测试评分工具(https://charactercalculator.com/flesch-reading-ease/)对 ChatGPT 和 Claude 的回答进行了评估[6]。可读性评估采用了两个不同的指标:Flesch 阅读容易度分数和 Flesch-Kincaid 年级水平。Flesch 可读性测试使用了考虑句子长度和单词数量的数学公式。Flesch 阅读容易度分数是一个介于 1 到 100 之间的数值,分数越高表示文本越易读。Flesch-Kincaid 年级水平分数表示理解文本所需的平均教育年数,分数越低表示可读性越高[7]。

1.3　可理解性和可操作性评估

患者教育材料评估工具(patient education materials assessment tool for print materials,PEMAT-P)是一种经过验证的标准化工具,用于评估患者信息的

可理解性和可操作性[8-9]。该工具包含24个问题，分为两部分：可理解性评估（问题1～17）和可操作性评估（问题18～24）。当不同背景和健康素养的个体能够理解信息并复述关键内容时，材料被视为可理解；当个体能够根据信息识别出可采取的行动时，材料被视为具有可操作性。在分析聊天机器人的回复前，两位评审者阅读了PEMAT-P用户指南。每个问题的评分标准为："0"表示不同意，"1"表示同意，"NA"表示不适用。各部分的得分相加后，除以该部分的总适用问题数，得出百分比评分，并最终合并为PEMAT总百分比评分。得分越高，表示材料的可理解性、可操作性或两者的综合性越高。每条回复的评分为两位评审者的平均分，这项评估由两名评审员独立完成，计算评分的平均值。

1.4 准确性评估

关于聊天机器人的一个问题在于，它们依赖于处理数十亿网络内容的庞大学习模型，通过语言建模生成回复，这些回复可能存在不准确、不完整或过时的风险。为了解决这些问题，本文评估了回复的准确性和全面性。其中，准确性定义为"接近真实值的程度"，而全面性定义为"覆盖广泛范围的完整性"[10]。这项评估由两名耳鼻咽喉头颈外科专家完成。两位专家分别使用5分Likert量表对每条回复进行独立评分，评分范围从1分（完全不准确/不全面）到5分（完全准确/全面）。专家们在评分时对聊天机器人的来源保持盲评状态。最终计算两位专家评分的平均值。

1.5 质量评估

使用扩展的基于证据的患者教育信息质量评估工具（evidence-based quality of information in patient education，EQIP）工具对质量进行二次评分[11]。该工具由36个项目（或标准）组成，用于评估患者教育文件的内容、识别和结构。在这项研究中，确定了19项标准与分析相关（内容数据：第1～11项；结构数据：第12～19项）。修改后的EQIP工具使用二元刻度"是（Yes）"与"否（No）"或"不适用"（NA）来避免歧义[11]，由第3位专家审核完成。根据这个修改后的EQIP工具审查了所有36个问题的每个ChatGPT和Claude回复，此外，通过统计每个来源获得"是"响应的标准数量来计算每个问题的得分。

1.6 统计学分析

使用R语言（4.4.1版本）完成统计学分析，*Mann-Whitney U*检验以比较两组之间的差异；χ^2检验比较不同组别在某分类变量上的分布差异，$P<0.05$为差异具有统计学意义。

2 结果

2.1 可读性

2.1.1 字数比较 整体来看，ChatGPT o1生成的答案明显更为详细，各类别回答长度均显著高于Claude 3.5 Sonnet，适合需要全面解释的医疗场景；而Claude 3.5 Sonnet回答简洁，更适合快速临床信息传递。①基础知识部分：ChatGPT o1为367±166.68，Claude 3.5 Sonnet为180.38±76.50，$U=51.5$，$P=0.0032$（图1A）。②诊断相关部分：ChatGPT o1为389±109.15，Claude 3.5 Sonnet为177±33.58，$U=8$，$P=0.01$（图1B）。③治疗相关部分：ChatGPT o1为390.2±169.29，Claude 3.5 Sonnet为203.73±77.27，$U=33$，$P=5.8\times10^{-4}$（图1C）。④总体字数对比：ChatGPT o1为381.56±153.28，Claude 3.5 Sonnet为189.36±69.09，$U=185$，$P=1.9\times10^{-7}$（图1D）。

2.1.2 平均阅读难易度分数比较 阅读难易度分析显示，ChatGPT o1答案的易读性显著高于Claude 3.5 Sonnet，其内容措辞更易被普通读者读懂，更适合临床信息普及与患者教育；而Claude 3.5 Sonnet表达更复杂，对读者医学背景要求更高。①基础知识部分：ChatGPT o1为16.35±13.24，Claude 3.5 Sonnet为4.66±8.82，$U=47.50$，$P=0.047$（图2A）。②诊断相关部分：ChatGPT o1为0±0，Claude 3.5 Sonnet为6.21±9.32，$U=16$，$P=0.032$（图2B）。③治疗相关部分：ChatGPT o1为9.39±8.66，Claude 3.5 Sonnet为0±0，$U=22.5$，$P=2.8\times10^{-5}$（图2C）。④总体平均阅读难易度分数：ChatGPT o1为11.20±11.16，Claude 3.5 Sonnet为1.68±5.64，$U=222$，$P=1.4\times10^{-7}$（图2D）。

2.1.3 年级分数比较 年级分数分析证实Claude 3.5 Sonnet内容复杂性显著高于ChatGPT o1，需要读者具备更高教育或专业背景才能理解。这表明ChatGPT o1文本对大众更友好，而Claude 3.5 Sonnet内容更适合具有较高医学知识水平的专业人群。①基础知识部分：ChatGPT o1为15.69±2.28，Claude 3.5 Sonnet为43.94±37.56，$U=1$，$P=2.1\times10^{-5}$（图3A）。②诊断相关部分：ChatGPT o1为17.47±1.67，Claude 3.5 Sonnet为50.10±11.58，$U=0$，$P=1.6\times10^{-4}$（图3B）。③治疗相关部分：ChatGPT o1为203.73±77.27，Claude 3.5 Sonnet为

67.02 ± 40.59, $U = 8$, $P = 8.1 \times 10^{-7}$(图 3C)。④总体平均年级分数:ChatGPT o1 为 16.70 ± 2.03,Claude 3.5 Sonnet 为 54.93 ± 35.81, $U = 7$, $P = 5.5 \times 10^{-13}$(图 3D)。

2.2 可理解性和可操作性

2.2.1 PEMAT-P 可理解性分数 PEMAT-P 评分分析显示,Claude 3.5 Sonnet 在总体可理解性上明显优于 ChatGPT o1,尤其在基础知识与诊断相关部分表现出显著优势。①基础知识部分:ChatGPT o1 为 0.58 ± 0.08,Claude 3.5 Sonnet 为 0.67 ± 0.13, $U = 44$, $P = 0.04$(图 4A)。②诊断相关部分:ChatGPT o1 为 0.57 ± 0.016,Claude 3.5 Sonnet 为 0.70 ± 0.15, $U = 10$, $P = 0.023$(图 4B)。③治疗相关部分:ChatGPT o1 为 0.58 ± 0.026,Claude 3.5 Sonnet 为 0.63 ± 0.096, $U = 72.5$, $P = 0.09$(图 4C)。④总体而言:ChatGPT o1 为 0.058 ± 0.049,Claude 3.5 Sonnet 为 0.46 ± 0.12, $U = 346.5$, $P < 0.05$(图 4D)。

2.2.2 PEMAT-P 可操作性分数 在可操作性方面,Claude 3.5 Sonnet 总体表现优于 ChatGPT o1,尤其在治疗相关内容上具有显著优势,而基础知识和诊断相关部分差异无统计学意义。①基础知识部分:ChatGPT o1 为 0.095 ± 0.13,Claude 3.5 Sonnet 为 0.087 ± 0.14, $U = 76.5$, $P = 0.66$(图 5A)。②诊断相关部分:ChatGPT o1 为 0.083 ± 0.044,Claude 3.5 Sonnet 为 0.28 ± 0.24, $U = 19$, $P = 0.15$(图 5B)。③治疗相关部分:ChatGPT o1 为 0.056 ± 0.055,Claude 3.5 Sonnet 为 0.38 ± 0.17, $U = 11$, $P < 0.05$(图 5C)。④总体方面:ChatGPT o1 为 0.076 ± 0.087,Claude 3.5 Sonnet 为 0.25 ± 0.22, $U = 372$, $P = 0.015$(图 5D)。

2.3 准确性

图 6 展示了 ChatGPT o1 与 Claude 3.5 Sonnet 在准确度方面的对比结果。ChatGPT o1 在信息准确性方面的表现明显优于 Claude 3.5 Sonnet。具体分析显示,两种 AI 模型在基础知识层面表现相近,差异无统计学意义。然而,在诊断与治疗相关内容

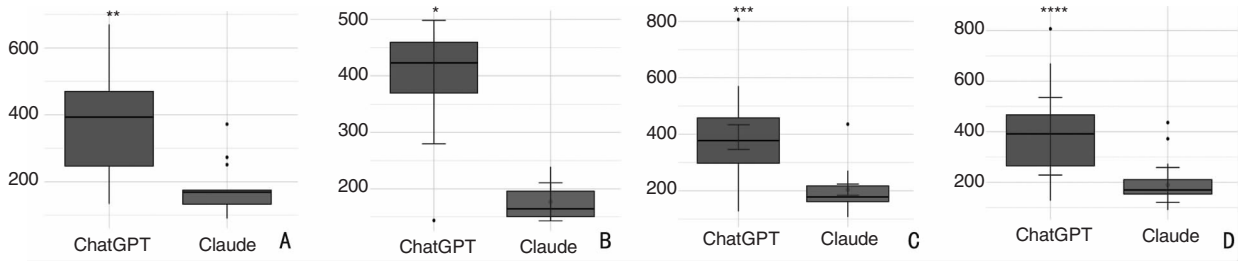

图 1 两种模型可读性字数对比 A:基础知识部分;B:诊断内容部分;C:治疗内容部分;D:总体字数分布

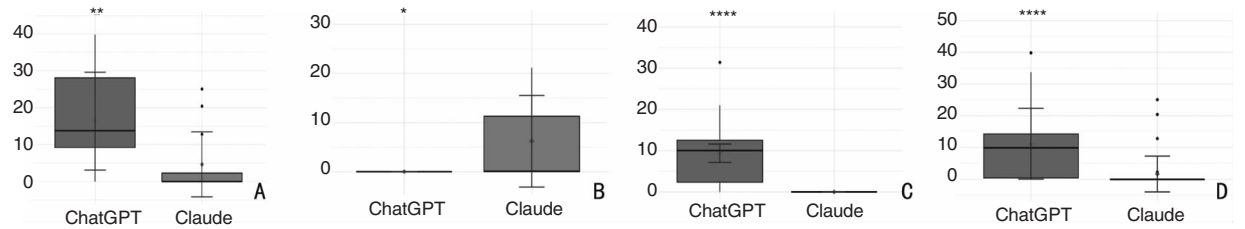

图 2 两种模型可读性阅读难易度对比 A:基础知识部分;B:诊断内容部分;C:治疗内容部分;D:总体阅读难易度分布

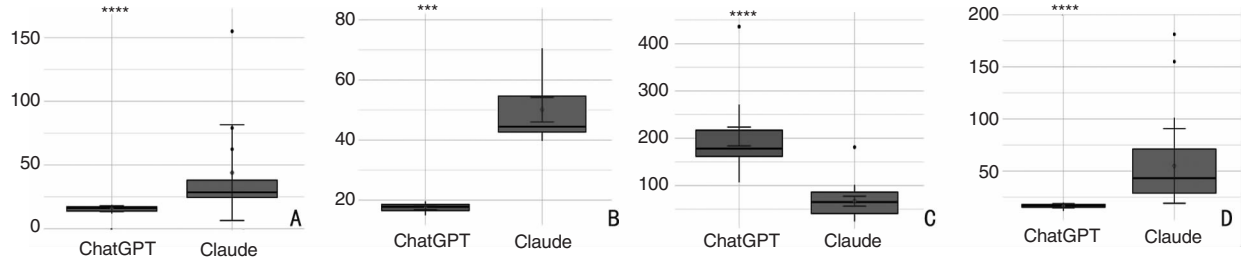

图 3 两种模型可读性年级分数对比 A:基础知识部分;B:诊断内容部分;C:治疗内容部分;D:总体年级分数分布

方面，ChatGPT o1 均表现出显著优势。①基础知识方面：ChatGPT o1 为 4.73 ± 0.39，Claude 3.5 Sonnet 为 4.69 ± 0.48，$U = 84$，$P = 1$（图 6A）。②诊断相关内容：ChatGPT o1 为 4.94 ± 0.18，Claude 3.5 Sonnet 为 4.63 ± 0.35，$U = 15.5$，$P = 0.049$（图 6B）。③治疗相关方面：ChatGPT o1 为 4.97 ± 0.13，Claude 3.5 Sonnet 为 4.60 ± 0.39，$U = 51$，$P = 0.002\ 3$（图 6C）。④总体准确度：ChatGPT o1 为 4.88 ± 0.28，Claude 3.5 Sonnet 为 4.58 ± 0.37，$U = 385.5$，$P = 0.002\ 2$（图 6D）。

2.4　质量

图 7 和表 1 展示了 ChatGPT o1 与 Claude 3.5 Sonnet 在修改后的 EQIP 标准下的对比结果。对于每项标准，我们分别统计了 ChatGPT o1 和 Claude 3.5 Sonnet 回答中获得“是”（Yes）响应的来源数量。结果显示，在满足各标准的平均来源数量方面，两者无显著统计学差异（14.21 ± 16.54 vs. 12.79 ± 16.10，$U = 177.5$，$P = 0.94$）（图 7A）。这一结果表明，从整体上来看，两种方法在提供符合 EQIP 标准的信息来源方面表现相近，未表现出显著的优劣之分。对其中 19 个单独标准的分析中（表 1），有 1 个标准在 ChatGPT o1 与 Claude 3.5 Sonnet 的回答间存在显著差异（$\chi^2 = 38.29$，$P < 0.05$），说明其在特定标准的表现可能存在差别。

此外，我们通过统计每个来源获得“是”响应的标准数量来计算每个问题的得分。图 8 结果显示，平均得分存在显著统计学差异，其中 ChatGPT o1 的平均得分（7.47 ± 1.28）显著高于 Claude 3.5 Sonnet 的平均得分（5.75 ± 1.20，$U = 140$，$P < 0.05$）（图 7B）。ChatGPT o1 的整体回答质量显著优于 Claude 3.5 Sonnet，提示其可能在信息完整性具有优势。

3　讨论

生成式 AI 技术正在全球范围内呈现快速发展之势。作为基于服务器的大规模语言模型，ChatGPT和Claude通过对海量互联网数据的深度学习，

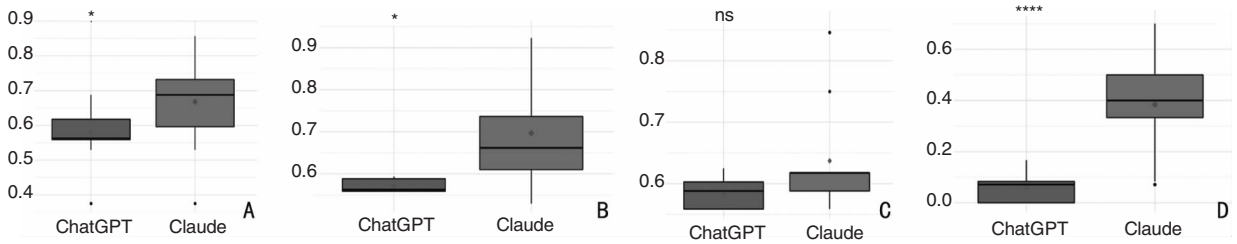

图 4　两种模型 PEMAT－P 可理解性对比　A：基础知识部分；B：诊断内容部分；C：治疗内容部分；D：PEMAT-P 整体可理解性评分　注：PEMAT-P（患者教育材料评估工具）。下同。

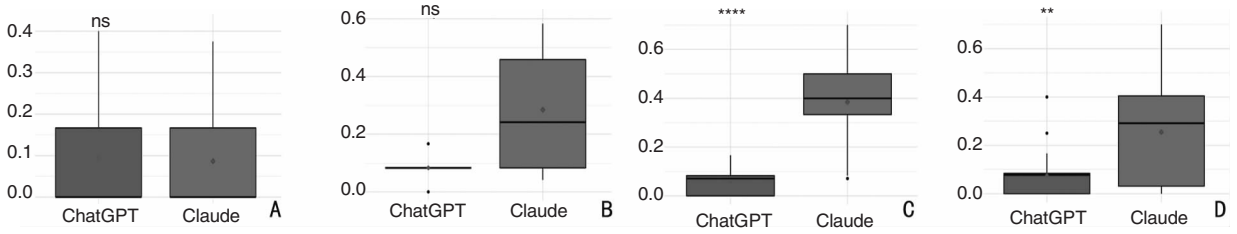

图 5　两种模型 PEMAT－P 可操作性对比　A：基础知识部分；B：诊断内容部分；C：治疗内容部分；D：PEMAT-P 整体可操作性评分

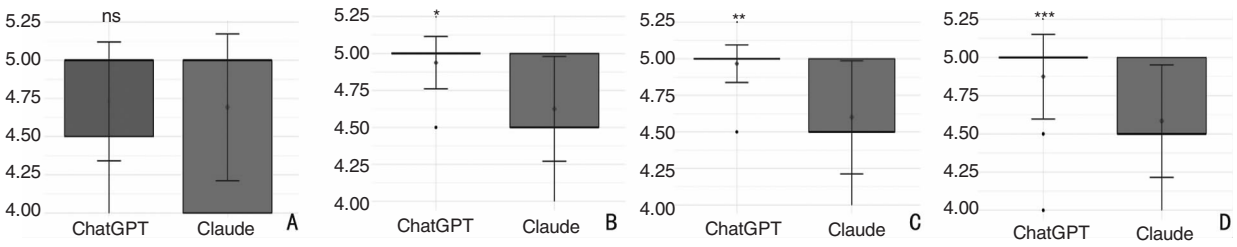

图 6　两种模型准确性对比　A：基础知识部分；B：诊断内容部分；C：治疗内容部分；D：总体准确性分布

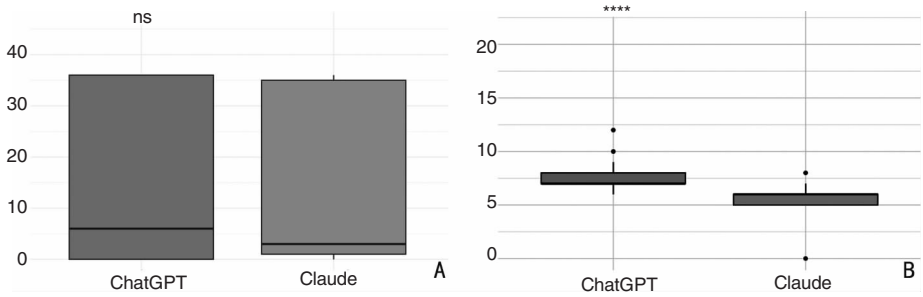

图 7　两种模型 EQIP 对比　A:符合 EQIP 标准的平均信息源数量;B:EQIP 总体评分分布　注:EQIP(患者教育信息质量评估工具)。下同。

表 1　ChatGPT 和 Claude 的质量评估所采用 EQIP 准则及各项标准的总体响应情况　(个)

标准	ChatGPT		Claude	
	是否包含	响应数量	是否包含	响应数量
内容数据				
要涵盖的主题的初步定义	Yes	36	Yes	35
	No	0	No	1
对先前定义主题的覆盖	Yes	36	Yes	35
	No	0	No	1
医疗干预目的定义*	Yes	29	Yes	2
	No	7	No	34
对接受者的量化益处描述	Yes	0	Yes	2
	No	36	No	34
定性风险和副作用的描述	Yes	7	Yes	7
	No	29	No	29
并发症处理方式的描述	Yes	1	Yes	1
	No	35	No	35
解决生活质量问题	Yes	0	Yes	2
	No	36	No	34
患者可采取的预防措施描述	Yes	0	Yes	0
	No	36	No	36
解决医疗干预费用和保险问题	Yes	0	Yes	0
	No	36	No	36
医院服务的具体联系方式	Yes	0	Yes	0
	No	36	No	36
其他可靠信息和支持渠道的详细说明	Yes	3	Yes	3
数据结构	No	33	No	33
使用日常语言,并对复杂词汇或专业术语进行解释	Yes	6	Yes	6
	No	30	No	30
对所有药物或产品均使用通用名称	Yes	7	Yes	6
	No	29	No	30
对读者的个性化称呼	Yes	1	Yes	1
	No	35	No	35
信息清晰(无歧义或矛盾)	Yes	36	Yes	36
	No	0	No	0
关于风险与收益的平衡信息	Yes	6	Yes	6
	No	30	No	30
按照逻辑顺序展示信息	Yes	36	Yes	36
	No	0	No	0
令人满意的设计和布局(不包括图形或图表)	Yes	36	Yes	36
	No	0	No	0
清晰且相关的图形或图表	Yes	0	Yes	0
	No	36	No	36

注:* 为差异具有统计学意义($P<0.05$);Yes(是);No(否)。

能够在实时对话中生成智能、流畅的回应,展现出强大的自然语言处理能力[12]。随着 ChatGPT 和 Claude 等大型语言模型向公众开放,医疗领域迎来新机遇。这些 AI 技术有望提升患者参与度、扩展医疗信息渠道,并有效降低医疗成本[13]。

本研究发现两种 AI 模型在答案特征上存在显著差异。ChatGPT o1 倾向于提供更详尽的回答。在阅读难易度评估中,Claude 3.5 Sonnet 的答案整体呈现更高的复杂度,可读性较低。这些发现表明,尽管 ChatGPT o1 提供更详细的答案,其措辞更容易接受,这可能使其更适合在临床实践中进行信息传播,易读的信息能有效提升患者与医护人员之间的沟通效率。

其次,从 PEMAT-P 评分体系评估的可理解性和可操作性来看 Claude 3.5 Sonnet 整体表现优于 ChatGPT o1。这一结果表明 Claude 生成的内容在实际应用场景中的指导性更强,可更有效地帮助患者或医护人员进行决策与操作。然而,Claude 的实际操作性仍存在提升空间,未来需要进一步完善其答案的实际可操作性,以更好地适应临床需求。

在准确性评估方面,虽然两种 AI 模型在基础知识领域表现相当,但 ChatGPT o1 在诊断相关内容和治疗相关方面均显示出显著优势。从总体准确度来看,ChatGPT o1 同样优于 Claude 3.5 Sonnet。这一

发现提示,尽管 Claude 3.5 Sonnet 在可理解性和可操作性方面具有优势,但在信息准确性方面仍有提升空间,特别是在临床诊疗相关内容上。

基于修改后的 EQIP 标准评估,虽然两种 AI 模型在满足各评估标准的平均来源数量上无显著差异,但在总体质量评分上,ChatGPT o1 的平均得分显著高于 Claude 3.5 Sonnet。值得注意的是,在19个评估标准中,仅有1个标准显示出显著差异。这些结果表明,尽管两个模型在大多数质量标准的达成度上相近,但 ChatGPT o1 在整体输出质量方面可能具有一定优势。医疗 AI 应用需具备高质量信息输出,以提升临床决策可信度和患者信任。持续优化 AI 内容质量是提高其临床适用性的关键方向。

虽然患者越来越依赖互联网获取信息,但缺乏有效辨识可靠信息的手段,人类专业知识与人工智能技术的整合正成为填补这一信息鸿沟的关键途径,为患者提供既准确又个性化的医疗内容[14]。

此外,本研究存在若干局限性:首先,研究仅针对头颈部单中心 Castleman 病相关的有限问题进行测试,尚无未验证 AI 在多中心型或其他耳鼻咽喉头颈外科中罕见病中的表现。其次,尽管本研究采用了标准化的评分体系,但在评估答案适宜性时仍不可避免地存在主观因素。此外,作为研究对象的 ChatGPT 和 Claude 等 AI 模型正处于持续迭代优化阶段,其性能可能会随版本更新而不断提升,这也使得当前研究结果具有时效性特征。

4 结论

本研究通过多维度评估比较了 ChatGPT o1 和 Claude 3.5 Sonnet 在头颈部单中心 Castleman 病相关问题上的表现。研究发现,Claude 3.5 Sonnet 倾向于提供更简洁的回答,并在可理解性和可操作性方面表现更优;而 ChatGPT o1 则在可读性、答案准确性和整体质量评分上具有显著优势。这些发现为临床实践中 AI 模型的选择和应用提供了参考依据,但仍需要更多研究来验证其在其他医学领域的表现。

参考文献:

[1] Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine[J]. Gastrointest Endosc, 2020, 92(4): 807-812.

[2] Castleman B, Iverson L, Menendez VP. Localized mediastinal lymph-node hyperplasia resembling thymoma[J]. Cancer, 1956, 9(4): 822-830.

[3] Wu D, Lim MS, Jaffe ES. Pathology of Castleman disease[J]. Hematol Oncol Clin North Am, 2018, 32(1): 37-52.

[4] 冯燚源,王安生,吴春成,等. 鼻腔 Castleman 病1例[J]. 中国耳鼻咽喉颅底外科杂志,2022,28(3):99-101.

[5] van Rhee F, Oksenhendler E, Srkalovic G, et al. International evidence-based consensus diagnostic and treatment guidelines for unicentric Castleman disease[J]. Blood Adv, 2020, 4(23): 6039-6050.

[6] Mnajjed L, Patel RJ. Assessment of ChatGPT generated educational material for head and neck surgery counseling[J]. Am J Otolaryngol, 2024, 45(5): 104410.

[7] Kasabwala K, Agarwal N, Hansberry DR, et al. Readability assessment of patient education materials from the American academy of otolaryngology head and neck surgery foundation[J]. Otolaryngol Head Neck Surg, 2012, 147(3): 466-471.

[8] Shoemaker SJ, Wolf MS, Brach C. Development of the patient education materials assessment tool (PEMAT): A new measure of understandability and actionability for print and audiovisual patient information[J]. Patient Educ Couns, 2014, 96(3): 395-403.

[9] Davis RJ, Ayo-Ajibola O, Lin ME, et al. Evaluation of oropharyngeal cancer information from revolutionary artificial intelligence chatbot[J]. Laryngoscope, 2024, 134(5): 2252-2257.

[10] Bellinger JR, Kwak MW, Ramos GA, et al. Quantitative comparison of chatbots on common rhinology pathologies[J]. Laryngoscope, 2024, 134(10): 4225-4231.

[11] Wei K, Fritz C, Rajasekaran K. Answering head and neck cancer questions: An assessment of ChatGPT responses[J]. Am J Otolaryngol, 2024, 45(1): 104085.

[12] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models[J]. PLoS Digit Health, 2023, 2(2): e0000198.

[13] Shen SA, Perez-Heydrich CA, Xie DX, et al. ChatGPT vs. web search for patient questions: What does ChatGPT do better? [J]. Eur Arch Otorhinolaryngol, 2024, 281(6): 3219-3225.

[14] Baker L, Wagner TH, Singer S, et al. Use of the internet and e-mail for health care information: Results from a national survey [J]. JAMA, 2003, 289(18): 2400-2406.

(收稿日期:2024-11-09;网络首发:2025-03-18)

本文引用格式:潘鑫,郜飞,陈亭亭,等. 颈部单中心型 Castleman 病临床辅助诊疗中 ChatGPT o1 和 Claude 3.5 Sonnet 的应用比较研究[J]. 中国耳鼻咽喉颅底外科杂志, 2025,31(6):76-82. DOI:10.11798/j. issn. 1007-1520. 202524552

Cite this article as: PAN Xin, GAO Fei, CHEN Tingting, et al. A comparative study on the application of ChatGPT o1 and Claude 3.5 Sonnet in the clinical adjuvant diagnosis and treatment of unicentric Castleman disease in the neck[J]. Chin J Otorhinolaryngol Skull Base Surg, 2025,31(6):76-82. DOI:10.11798/j. issn. 1007-1520. 202524552