·信息化护理·
·论　　著·

GPT-4 大语言模型对护理知识理解的测试研究

徐文博[1]，陈凤敏[1]，王超[2]，陈洁[3]，侯辉[3]

摘要：目的 探讨 GPT-4 大语言模型在护理教育中的应用潜力。**方法** 选用 GPT-4 对主管护师考试真题进行量化测试，并对答案准确率进行分类评价。**结果** GPT-4 的整体准确率为 81.00%。在知识记忆和简单选项题目上准确率较高，分别为 82.64%和 82.52%；在解答知识应用和复杂选项题目时，GPT-4 的准确率较低，分别为 76.60%和 70.97%。**结论** GPT-4 展现出作为护理教学和临床护理辅助工具的巨大潜力。未来研究应探索如何将大语言模型与外部知识源结合并创新应用方法，提升大模型生成内容的准确性。同时，护理教育工作者还应积极探索大模型提升学生自学能力和自我评价能力的方法。

关键词：大语言模型；　人工智能；　GPT-4；　护理教育；　护理教学；　护理知识；　试题；　测试
中图分类号：G642.0；TP181　**DOI：**10.3870/j.issn.1001-4152.2024.19.093

Testing research of GPT-4 large language model on nursing knowledge comprehension

Xu Wenbo, Chen Fengmin, Wang Chao, Chen Jie, Hou Hui. Department of General surgery, The First Affiliated Hospital of Jinzhou Medical University, Jinzhou 121000, China

Abstract: Objective To explore the application potential of GPT-4 large language model in nursing education. **Methods** GPT-4 was used to quantitatively test the Supervisory Nurse Examination, and the accuracy of the answers was classified and evaluated. **Results** The overall accuracy of GPT-4 was 81.00%. The accuracy of knowledge memorization and simple choice questions was 82.64% and 82.52%, respectively. The accuracy of GPT-4 was 76.60% and 70.97% respectively when solving knowledge application and complex choice questions. **Conclusion** GPT-4 shows great potential as an auxiliary tool for nursing teaching and clinical nursing. Future research should explore how to combine large language models with external knowledge sources and innovate application methods to improve the accuracy of content generated by large models. At the same time, nursing educators should actively explore ways to improve students' self-study ability and self-evaluation ability with large language models.

Keywords: large language model; artificial intelligence; GPT-4; nursing education; nursing teaching; nursing knowledge; test questions; test

自 2022 年 12 月 ChatGPT 大语言模型诞生以来，其在各个领域的应用受到广泛关注与热议。2023 年 3 月，OpenAI 公司推出了 GPT-4，与之前的 GPT-3.5 相比，GPT-4 在多个方面有了显著提升。有媒体报道，ChatGPT 在国外各类资格考试中得分可轻易赶超 90%的考生，而且其生成答案的速度遥遥领先[1]。目前，国内外关于大语言模型在护理教育领域的应用研究仍相对不足。2024 年 2 月检索 Scopus、PubMed、中国知网等国内外数据库显示，关键词包含“ChatGPT”的研究文献多达 5 300 篇以上，其中与护理教育相关的研究论文仅有 20 多篇，而且大多是综述类文献，探讨大语言模型运用于护理教育的实证研究较少。尽管如此，大部分学者对人工智能大语言模型应用于护理教育领域的潜力持积极态度，一致认为大语言模型将推动护理教学与考核的变革[2-5]。GPT-4 作为基于 Transformer 架构和自监督预训练的大语言模型，是目前 OpenAI 发布的最先进、最强大的语言模型[6]。GPT-4 大语言模型较于其他人工智能大语言模型，具有强大的自然语言处理能力和更高的准确性并支持多模态输入，其应用场景更广泛。因此，本研究选用 GPT-4 作为测试对象，拟通过使 GPT-4 完成试卷考核的方式来对其进行量化测试，并探讨其在护理教育中的应用潜力和风险，为我国护理教育领域应用 ChatGPT 大语言模型提供参考。报告如下。

1　资料与方法

1.1　一般资料　选取 2021 年全国卫生专业技术护理学专业（主管护师）资格考试真题（下称“主管护师考试”）为测试试卷。主管护师考试科目共 4 门，包括基础知识、相关专业知识、专业知识、专业实践能力，每份试卷 100 题，共 400 题。均为单项选择题，每题包含 5 个选项，每科满分 100 分。全国统考真题具有一定权威性与规范性，其内容涵盖内科、外科、妇科、儿科、社区护理、护理管理等试题，临床实践性强、知识点覆盖全面、题型多样、难度适中，注重护理人员综

作者单位：1.锦州医科大学附属第一医院普外科（辽宁 锦州，121000）；2.辽宁工业大学图书馆；3.桂林理工大学图书馆
徐文博：女，硕士，副主任护师，ruwu79@126.com
通信作者：陈凤敏，chenfengmin198706@163.com
科研项目：2023 年度教育部人文社会科学研究规划基金（23YJA870002）
收稿：2024-05-30；修回：2024-07-22

合能力考核。因此，该试题比较适合用于考察 GPT-4 大语言模型对护理学知识的掌握程度与运用能力。

1.2　方法

1.2.1　试题预处理　邀请本校护理学院 2 名护理教学年限≥10 年的护理学专家对 400 道真题进行分类。1 名专家为正高级职称、另 1 名专家为副高级职称，2 人均为硕士生导师。2 名专家分别对 400 道真题进行分类，如遇分歧则与第 3 名专家进行协商决定。分类方法：①题干信息分类。按照题干信息依据布鲁姆认知水平评估理论[7]将试题为知识记忆、知识理解、知识应用、临床场景 4 个类别。②题型分类。按照题型分为病例题和非病例题。③难易程度分类。按照题目内容的难易程度分为简单选项题、一般选项题和复杂选项题。分类原则[8-9]：题目选项字数的平均数为 39 字，选项字数为 39 字以下为简单选项题，选项字数为 40～78 字为一般选项题，选项字数为 79 字以上为复杂选项题。④问题要求分类。按照问题要求分为积极选项题和消极选项题。分类原则：如果问题要求选出正确或肯定的答案为“积极选项题”，如果问题需要识别不正确、不包括、不妥当等否定的答案则为“消极选项题”。题型分类见表 1。

表 1　2021 年主管护师资格考试题型分类（n=400）

题

分类	题型类别	科目				合计
		基础知识（n=100）	相关专业知识（n=100）	专业知识（n=100）	专业实践（n=100）	
题干信息	知识记忆	59	69	71	66	265
	知识理解	15	13	8	13	49
	知识应用	13	13	10	11	47
	临床场景	13	5	11	10	39
题型	非病例	86	86	86	79	337
	病例	14	14	14	21	63
难易程度	简单选项	82	71	79	77	309
	一般选项	9	18	16	17	60
	复杂选项	9	11	5	6	31
问题要求	积极选项	90	87	85	82	344
	消极选项	10	13	15	18	56

1.2.2　对话语言输入方法及相关调试流程　经过反复对比实验，最终确定以下提示语作为 GPT-4 的输入（Prompt）：“你是一位护理教育领域的专家，请针对我提供的护理试题，给出正确答案，并详细解释你选择该答案的理由，你的答案需要准确无误，同时理由也要充分，使得普通人也能易于理解。”为了避免“记忆效应”且符合国家法律法规，本研究通过调用应用程序编程接口（Application Programming Interface，API）的方法完成测试，GPT-4 在 OpenAI 的模型代号为“gpt-4.0-turbo”，可通过调用 API 进行使用，具体操作方法包括①安装 OpenAI 依赖包，输入 API Key；②确定模型、Prompt 及问题 model＝“gpt-4.0-turbo”，messages＝[{“role”：“system”，“content”：“护理试题解答，根据提供的试题及选项选择正确的答案，并给出详细解释”}，{“role”：“user”，“content”：全国卫生专业技术护理学专业资格考试题目和提示语}]。③调用 API 接口，生成题目答案选项和解释，将其拼接到已有的标准答案的下一列，保存到 Excel 表格中。④重复后面 2 个操作 400 次，生成 400 组答案及解析。

1.2.3　答案统计与计算方法　①如果由 GPT-4 生成的答案与参考答案一致，则判断为该题项回答正确。②由 2 名护理学专家对大模型输出的答案进行人工审核，如遇分歧，与第 3 名护理学专家进行协商决定。审核方式为：答案正确且与解析理由一致视为审核通过；答案与解析结果不一致则视为审核不通过，包括答案正确但解析错误、答案错误但解析正确 2 种情况。专家审核通过率为 GPT-4 答题最终准确率。③最后统计得出试题各部分测验结果。采用频数和构成进行统计描述。

2　结果

2.1　GPT-4 生成试卷各科目答案正确率及专家审核通过率　见表 2。

表 2　GPT-4 生成试卷各科目答案正确率及专家审核通过率

题（%）

项目	答案正确	答案正确解析错误	专家审核通过（答案准确）
基础知识（n=100）	86(86.00)	2(2.00)	84(84.00)
相关专业知识（n=100）	79(79.00)	7(7.00)	72(72.00)
专业知识（n=100）	88(88.00)	2(2.00)	86(86.00)
专业实践（n=100）	85(85.00)	3(3.00)	82(82.00)
合计	338(84.50)	14(3.50)	324(81.00)

2.2　GPT-4 生成试卷不同类型试题答案情况及专家审核通过率　见表 3。

表 3　GPT-4 生成试卷不同类型试题答案情况及专家审核通过率　　题（%）

项目	题干信息分类				难易程度分类			题型分类		问题要求分类	
	知识记忆（n=265）	知识理解（n=49）	知识应用（n=47）	临床场景（n=39）	简单选项（n=309）	一般选项（n=60）	复杂选项（n=31）	非病例题（n=337）	病例题（n=63）	积极选项（n=344）	消极选项（n=56）
答案正确	230(86.79)	40(81.63)	37(78.72)	31(79.49)	268(86.73)	48(80.00)	22(70.97)	286(84.87)	52(82.54)	289(84.01)	49(87.50)
答案错误	35(13.21)	9(18.37)	10(21.28)	8(20.51)	41(13.27)	12(20.00)	9(29.03)	51(15.13)	11(17.46)	55(15.99)	7(12.50)
答案解析不一致	16(6.04)	3(6.12)	2(4.26)	2(5.13)	20(6.47)	2(3.33)	1(3.23)	19(5.64)	4(6.35)	17(4.94)	6(10.71)
答案正确解析错误	11(4.15)	1(2.04)	1(2.13)	1(2.56)	13(4.21)	1(1.67)	0(0)	12(3.56)	2(3.17)	9(2.62)	5(8.93)
答案错误解析正确	5(1.89)	2(4.08)	1(2.13)	1(2.56)	7(2.27)	1(1.67)	1(3.23)	7(2.08)	2(3.17)	8(2.33)	1(1.78)
专家审核通过（答案准确）	219(82.64)	39(79.59)	36(76.60)	30(76.92)	255(82.52)	47(78.33)	22(70.97)	274(81.31)	50(79.37)	280(81.40)	44(78.57)

3　讨论

3.1　GPT-4 对护理知识的理解能力　本研究显示，GPT-4 考试试卷答题正确率为 84.50%，整体表现比较优异，但经过护理学专家的人工审核后，其准确率下降

至 81.00%。GPT-4 在各科目试卷的准确率：基础知识、相关专业知识、专业实践能力高于平均值 81.00%，相关专业知识的准确率相对较低为 72.00%。从题型分类看，GPT-4 大语言模型在知识记忆类题目上表现最好，经护理学专家人工审核后，准确率为 82.64%，在处理简单选项类题目时也显示出较高的准确率为 82.52%，这表明 GPT-4 能够基本理解护理学的基本概念、原则和方法。然而在处理复杂选项类题目时，其准确率下降至 70.97%，这表明 GPT-4 在处理需要复杂临床推理或多步骤解决方案的病例时，可能会遇到一些挑战。这在知识应用类题目（准确率 76.60%）和临床场景类题目（准确率 76.92%）测试中也有所体现，上述结果与穆兰等[10]的研究结论基本一致。总体而言，GPT-4 在主管护师资格考试真题测试中表现出了较强的理论知识掌握能力和基本问题解决技能，但是在处理高复杂度的临床任务时仍显示出一定的不足。此外，GPT-4 生成的答案与其给出的解析之间存在逻辑不一致或明显错误的情况。在 400 道真题测试中共有 23 题（5.75%）出现了答案与解析不一致的情况，其中答案正确但解析错误 14 题（3.50%），答案错误但其解析正确 9 题（2.25%）。答案与其解析不一致的题目涵盖了所有题目类型并无明显的规律。这与相关研究[11-12]的结论类似：大语言模型在生成答案和解析时偶尔会出错，而且错误具有随机性和偶然性。

3.2 GPT-4 在主管护师考试真题测试中的局限性

GPT-4 共有 62 题（15.50%）的答案出现错误，其中 23 题（5.75%）出现了答案与解析不一致的情况。从错误题型分布看，GPT-4 在知识应用、临床场景、复杂选项、知识理解出现错误的概率较高。研究者认为导致大模型生成错误答案的原因可归纳为如下几类：一是逻辑推理错误，主要集中在知识理解题和临床场景题上。如“某儿科病房于 2015 年 10 月 3—10 日共收治患儿 60 例，其中新生儿病房 15 例，有 3 例发生轮状病毒感染，新生儿轮状病毒感染率是？”GPT-4 认为是 5%，可见大模型推理计算过程出现了明显的偏差。二是对题干信息的理解偏差，主要集中于复杂选项题或知识应用题。如“在护理人才的智能结构中，属于能力结构的是？”GPT-4 认为答案是“实践能力”，很显然 GPT-4 没有理解“智能结构”这一概念的含义。再如“结肠癌 Dukes 分期中，C 期的描述最准确的是？”题干信息强调“最准确”，大模型可能理解为描述正确，所以答案出现错误。三是排除干扰信息能力不足，主要集中于知识应用或复杂选项题。如“患者张某 36 岁，肾移植术后 36 h，出现少尿、血肌酐持续升高，并伴有高热、寒战，该患者可能出现了？”很显然该患者出现了排斥反应，GPT-4 可能受到干扰选项信息的误导最终选择了错误答案。此外，在某些情况下，GPT-4 偶尔出现了基本常识与科学做法混淆的现象。如“原发性血小板减少性紫癜患儿发生鼻出血时，应首先采取的措施是？”GPT-4 选择“用干燥的纱布压迫止血”，这确实符合基本常识，但科学的做法是“用浸麻黄碱的纱条压迫止血”。GPT-4 大模型在护理学真题测试中展现强大的能力，但也暴露出在记忆、推理、信息理解、干扰排除以及专业知识应用等方面的不足。

3.3 GPT-4 运用于护理教育领域的前景与注意事项

传统的护理教学仍是以教科书为基础结合多媒体技术的教学模式，教科书经过逐层审核和多年应用其知识信息虽然准确无误，但对学生批判性思维的培养和启发不足。GPT-4 具有实时反馈和自适应学习的能力，这对培养学生的批判性思维和知识应用将发挥重要作用。目前，GPT-4 的训练数据来源并不透明，据了解，GPT-4 并没有纳入中国知网、PubMed 和 Web of Science 等数据库，这可能导致其生成的内容在准确性和可靠性上存在不足，但 GPT-4 仍可作为护理教学和培训的辅助工具[13]。未来研究可将大语言模型与外部知识源或自建知识库相结合，进一步提高其准确性[14]。GPT-4 为师生提供了一个强大的知识库和教学辅助工具[15]，护理学生可以通过与 GPT-4 的交互快速获取准确的护理学知识，提高学习效率。需要注意的是，医学知识始终处于快速更新中，GPT-4 可能无法实时反映最新的研究进展或临床指南，会影响其在教学或临床实践中的可用性，在确保数据准确性的同时，保持大模型的训练数据不断更新至关重要[16]。此外，GPT-4 在消极选项类型题目上的高准确率表明了其良好的错误排除能力，可将 GPT-4 应用于临床护理中辅助临床护理决策，以提高护理质量和效率。但 GPT-4 在面对复杂选项类题目时其准确率显著下降，反映出 GPT-4 在临床推理和多步骤解决方案上的能力还有待提升，提示可通过对 GPT-4 进行更深入地训练和优化，特别是加强其在临床护理场景中的实践应用训练，以提高其临床推理和问题解决能力[17]。大语言模型应用于临床可能会触及个人健康信息，这需要医护人员严格遵守相关的医疗信息隐私保护法规。

4 结论

本研究以目前公认的最先进、最强大的 GPT-4 大语言模型为量化测试对象，在一定程度上揭示了人工智能大语言模型在护理教育与临床护理中应用的优势与局限。未来的研究应聚焦于提升大语言模型生成内容的准确性，并创新应用方法以减少错误信息的风险。本研究仅 400 道真题作为测试对象，样本量较小，可能无法全面展示大语言模型的能力，且仅使用 GPT-4 作为唯一测试大语言模型，可能无法全面反映大语言模型的能力。未来研究需纳入多个大语言模型进行综合测试，并进行性能比较，进一步推动

大语言模型在护理领域的深入应用与发展。

参考文献：

[1] 钱童心. ChatGPT 已能通过高难度考试 未来教育走向何方？[N]. 第一财经日报，2023-02-14(A04).

[2] 李钥，淮盼盼，杨辉. ChatGPT 在护理教育中的应用状况及优劣分析[J]. 护理学杂志，2023，38(21)：117-121.

[3] 马武仁，弓孟春，戴辉，等. 以 ChatGPT 为代表的大语言模型在临床医学中的应用综述[J]. 医学信息学杂志，2023，44(7)：9-17.

[4] 郭彩霞，郭彩旭，史晓宁，等. ChatGPT 赋能护理实践：前景、风险及对策[J]. 协和医学杂志，2023，14(6)：1170-1174.

[5] 罗华宇，许敏，曾朝蓉，等. ChatGPT 在护理领域中应用的前景与挑战[J]. 中华护理教育，2023，20(12)：1520-1523.

[6] Wu T，He S，Liu J，et al. A brief overview of ChatGPt：the history，status quo and potential future development[J]. IEEE/CAA Journal of Automatica Sinica，2023，10(5)：1122-1136.

[7] 王贤勇，喻斌. 基于布鲁姆学习目标和 SOLO 分类理论的试题对比评析[J]. 物理教师，2023，44(1)：84-88，92.

[8] 杜庆贤. ChatGPT 在财会工作中的应用探索：基于文心一言、ChatGPT、ChatSonic 测试[J]. 会计之友，2024(4)：130-137.

[9] Taira K，Itaya T，Hanada A，et al. Performance of the large language model ChatGPT on the National Nurse Examinations in Japan：evaluation study[J]. JMIR Nurs，2023，6：e47305.

[10] 穆兰，徐文博，王学通. 类 ChatGPT 大语言模型在护理教育中应用的实证探讨与前景展望[J/OL]. 卫生职业教育，2024，42：1-4[2024-07-19]. http://kns.cnki.net/kcms/detail/62.1167.R.20240528.1830.002.html.

[11] 王超，孔祥辉. 大型预训练语言模型在网络健康信息鉴别中的应用探讨[J]. 农业图书情报学报，2023，35(6)：51-59.

[12] Thirunavukarasu A J，Hassan R，Mahmood S，et al. Trialling a large language model (ChatGPT) in general practice with the applied knowledge test：observational study demonstrating opportunities and limitations in primary care[J]. JMIR Med Educ，2023，9：e46599.

[13] Elango A，Kannan N，Anandan I，et al. Testing the knowledge and interpretation skills of ChatGPT in pharmacology examination of phase Ⅱ MBBS[J]. Indian J Pharmacol，2023，55(4)：266-267.

[14] Branum C，Schiavenato M. Can ChatGPT accurately answer a PICOT question? Assessing AI response to a clinical question[J]. Nurse Educ，2023，48(5)：231-232.

[15] Huang H. Performance of ChatGPT on registered nurse license exam in Taiwan：a descriptive study[J]. Healthcare(Basel)，2023，11(21)：2855.

[16] Hirosawa T，Harada Y，Yokose M，et al. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 Chatbot for clinical vignettes with common chief complaints：a pilot study[J]. Int J Environ Res Public Health，2023，20(4)：3378.

[17] 刘乾坤，戴婧佼，庞佳雪，等. ChatGPT 技术在护理教育中的展望[J]. 护士进修杂志，2024，39(12)：1285-1290.

（本文编辑　黄辉，吴红艳）

（上接第 80 页）

[3] Zhang J，Yan Q Y，Yue S T. Nursing research capacity and its management in China：a systematic review[J]. J Nurs Manag，2020，28(2)：199-208.

[4] 李臣之，阮沁汐，纪海吉. 研究生学习获得感影响因素的质性探究[J]. 现代教育管理，2020，11：102-110.

[5] Forgrave D. Gibbs′s reflective cycle[EB/OL]. (2020-11-11)[2024-03-10]. https://www.ed.ac.uk/reflection/reflectors-toolkit/reflecting-on-experience/gibbs-reflective cycle.

[6] Li Y F，Chen W J，Liu C Q，et al. Nurses′ psychological feelings about the application of Gibbs reflective cycle of adverse Events[J]. Am J Nurs Sci，2020，9(2)：74.

[7] 杜静，徐明明，廖国琼，等. 护生基于 Gibbs 反思循环圈撰写实习反思日志的效果[J]. 护理学杂志，2021，36(24)：65-68.

[8] Grant A M，Franklin J，Langford P. The self-reflection and in sight scale：a new measure of private self-consciousness[J]. Soc Behav Pers，2002，30(8)：821-835.

[9] 刘健，刘露，陈秀红，等. 自我反思与洞察力量表中文版在精神障碍患者中应用的效度和信度[J]. 中国心理卫生杂志，2018，32(5)：369-374.

[10] Chen F F，Chen S Y，Pai H C. Self-reflection and critical thinking：influence of professional qualifications on registered nurses[J]. Contemp Nurse，2019，55(1)：59-70.

[11] Kim J H，Shin H S. Effects of self-reflection-focused career course on career search efficacy，career maturity，and career adaptability in nursing students：a mixed methods study[J]. J Prof Nurs，2020，36(5)：395-403.

[12] 颜红波，于红静，关玉仙，等. 反思日记联合案例法在神经外科低年资护士培训中的应用[J]. 护理学杂志，2019，34(3)：93-95.

[13] 杨敏菲，李卫东，胡佩欣. 护理学生学业复原力与自我控制的相关性研究[J]. 中国校医，2019，33(3)：188-190.

[14] 李海秋. 高师院校学前教育专业学生反思现状的调查[J]. 教育观察，2019，8(40)：136-137.

[15] Ardian P，Hariyati R T S，Afifah E. Correlation between implementation case reflection discussion based on the Graham Gibbs Cycle and nurses′ critical thinking skills[J]. Enfermería Clínica，2019，29：588-593.

[16] 赵廷芬，张莉，张耕，等. 导师指导风格对护理硕士研究生学习获得感的影响研究[J]. 中华护理育，2023，20(8)：936-940.

（本文编辑　黄辉，吴红艳）