Advances in Clinical Medicine 临床医学进展, 2026, 16(6), 2655-2665
Published Online June 2026 in Hans. https://www.hanspub.org/journal/acm
https://doi.org/10.12677/acm.2026.1662488

Hans汉斯

主流大语言模型在输液港健康科普中的表现评估：可读性与质量分析

黄俊豪*，刘艳玲，杨永刚，杨　猛，叶文彬，赖雁玲#

中山大学肿瘤防治中心手术麻醉科，广东　广州

收稿日期：2026年5月29日；录用日期：2026年6月23日；发布日期：2026年6月30日

摘　要

目的：系统评估主流大型语言模型(LLMs)生成植入式静脉输液港(IVAPs)科普文本的可读性与专业质量，为临床健康教育工具的选择提供循证依据。方法：选取5款主流通用大模型(GPT-5、豆包、深度求索、通义千问、文心一言)，针对输液港5大核心主题生成100篇科普文本。采用7项国际通用可读性指标，结合中文版患者教育材料评估工具(C-PEMAT-P)和全球质量量表(GQS)，对文本进行多维度量化分析。结果：不同模型生成文本的专业质量存在极显著差异($P < 0.001$)，GPT-5综合表现最优，豆包和深度求索紧随其后；各模型文本的可读性同样存在显著差异。科普主题的复杂度会显著影响文本可读性，但不会改变高质量模型的内容产出水准，证实可读性与质量是两个相对独立的评价维度。结论：模型类型是决定输液港科普文本质量的核心因素，临床应用优先推荐GPT-5、豆包和深度求索。医护人员应优选高质量模型生成专业内容，再结合患者认知特点进行针对性的可读性优化，实现医学专业性与大众易懂性的平衡。

关键词

大型语言模型，植入式静脉输液港，健康科普，可读性，文本质量

Evaluation of Mainstream Large Language Models in Health Education on Infusion Port Care: Readability and Quality Analysis

Junhao Huang*, Yanling Liu, Yonggang Yang, Meng Yang, Wenbin Ye, Yanling Lai#

Department of Surgical Anesthesiology, Sun Yat-sen University Cancer Center, Guangzhou Guangdong

Received: May 29, 2026; accepted: June 23, 2026; published: June 30, 2026

*第一作者。
#通讯作者。

文章引用: 黄俊豪, 刘艳玲, 杨永刚, 杨猛, 叶文彬, 赖雁玲. 主流大语言模型在输液港健康科普中的表现评估: 可读性与质量分析[J]. 临床医学进展, 2026, 16(6): 2655-2665. DOI: 10.12677/acm.2026.1662488

Abstract

Objective: To systematically evaluate the readability and professional quality of health education texts on implantable venous access ports (IVAPs) generated by mainstream large language models (LLMs), thereby providing an evidence-based basis for selecting clinical health education tools. Methods: A total of 100 texts were generated by 5 widely available LLMs (GPT-5, Doubao, DeepSeek, Tongyi Qianwen, Wenxin Yiyan), covering 5 core education themes related to IVAPs. Seven internationally validated readability indices, the Chinese version of the Patient Education Materials Assessment Tool (C-PEMAT-P), and the Global Quality Scale (GQS) were used for multidimensional quantitative analysis of the texts. Results: Significant differences in text quality were found among the models ($P < 0.001$), with GPT-5 achieving the highest overall performance, followed by Doubao and DeepSeek; significant differences in text readability were also observed across models. Topic complexity significantly affected readability but not the quality of content produced by high-performing models. Topic complexity significantly affected readability but not quality, indicating that readability and quality are relatively independent evaluation dimensions. Conclusion: Model type is the key determinant of text quality on IVAPs, with GPT-5, Doubao, and DeepSeek being the optimal choices. Healthcare providers should prioritize high-quality models for professional content generation, then optimize readability based on patients' cognitive characteristics to balance medical accuracy and public comprehensibility.

Keywords

Large Language Models, Implantable Venous Access Port, Health Science Communication, Readability, Text Quality

1. 引言

植入式静脉输液港(Implantable Venous Access Ports, IVAPs)是一种完全埋植于皮下的中心静脉通路装置，目前已广泛应用于需长期接受化学治疗、肠外营养支持及频繁静脉穿刺的患者群体[1]。多项临床研究证实，相较于经外周静脉置入中心静脉导管(PICCs)，IVAPs 在长期留置中展现出显著的临床优势，不仅导管相关性血栓和局部感染的发生率更低，还因隐蔽性好、维护周期长，能极大程度地减轻患者的心理负担，显著提升患者的整体生活质量[2]。然而，IVAPs 的长期安全使用高度依赖于规范的日常管理；若在治疗间歇期缺乏标准化的冲管与封管维护，极易诱发导管功能障碍、继发性导管堵塞及血栓形成等严重并发症，这不仅会缩短装置的使用寿命，亦会加重患者的生理痛苦与医疗负担[3]。

慢性病与长期照护管理中，传统的健康教育模式常因临床时间紧缩及信息传递的主观差异性而受限。正如 Paterick 等[4]在 2017 年指出的，高质量的医患沟通需要充足的时间投入，但现实中临床时间的稀缺使得深度医患沟通难以实现；同时，传统口头宣教中可能存在的专业术语滥用、沟通方式不统一等问题，也会导致信息传递效果参差不齐，因此需要医生通过标准化沟通方法提升信息传递的准确性。随着互联网的普及，患者愈发依赖网络平台获取医疗资讯。然而，既往针对网络健康信息的系统综述表明，现有在线医学科普材料的整体可读性往往超出公众的平均健康素养水平，导致专业信息难以被有效转化与吸收[5]。近年来，自然语言处理技术的突破使得 LLMs 成为重塑医疗科普生态的有力工具。以生成式预训练为核心的大模型范式，Radford 等验证了通用语言模型的强泛化能力与长文本理解能力，为医疗知识的

通俗化转译、个性化科普交互提供了关键技术支撑[6]。以 ChatGPT 为代表的主流模型在逻辑推理与复杂文本生成方面展现出显著优势[7]，其在生物医学信息检索及临床问答中的应用价值已得到多项系统综述的充分肯定[8]。尽管如此，当前关于 LLMs 医疗科普生成质量的评估多聚焦于糖尿病[8]、心血管疾病[9]等常见慢性病。针对完全 TIVAP 这类专业性极强，且需向患者精准传达手术植入流程与日常维护规范的特殊医疗装置，LLMs 生成内容的专业质量与患者可读性之间的平衡关系尚缺乏系统评估。鉴于此，本研究旨在填补这一研究空白，通过多维度评估筛选出最优模型，并为未来基于 LLMs 生成高质量医疗科普材料提供循证策略。

2. 材料与方法

2.1. 伦理考量

本研究采用大型语言模型生成的文本数据，不涉及人体受试者、动物实验及个人健康信息。

2.2. 研究设计与数据收集

2.2.1. 问题设计

由 2 名拥有 5 年以上肿瘤护理及输液港管理经验的临床护士，共同设计 20 个患者高频咨询问题，涵盖 5 个核心健康教育主题(每个主题 4 个问题)。问题经 1 名资深肿瘤医师验证，确保其相关性、全面性及与患者需求的契合度(表 S1)。

2.2.2. 模型选择

纳入 5 款目前可广泛获取的主流通用大模型：GPT-5、豆包、深度求索、通义千问、文心一言。所有模型均使用免费版本，未进行自定义参数调整，以模拟真实应用场景。

2.2.3. 文本收集

数据收集于 2025 年 10 月进行。将每个问题依次输入各模型，生成的文本以纯文本格式记录(排除图片、超链接及格式元素)。共收集 100 篇文本(每模型 20 篇)。

2.3. 评估指标

2.3.1. 可读性评估

根据可读性经典原则[10]，采用 7 项已被广泛验证适用于医学文本分析的客观指标，从不同维度评估文本复杂度[11]：自动可读性指数(ARI)、弗莱士阅读 ease 分值(FRES)、冈宁迷雾指数(GFOG)、弗莱士－金凯德年级水平(FKGL)、科尔曼－廖指数(CL)、简单晦涩度测量(SMOG)、林塞尔写作指数(LW)。

2.3.2. 质量评估

C-PEMAT-P：中文版患者教育材料评估工具，专门用于评估印刷类患者教育材料的“可理解性”和“可操作性”，满分 24 分，是目前国际公认的患者健康材料评估金标准[12]。

GQS (Global Quality Scale)：5 分李克特量表，广泛用于评估互联网健康信息的整体质量与流畅度[13]，1 分 = 质量差，5 分 = 优秀。由 2 名临床护士独立评分。采用科恩卡帕系数(Cohen’s Kappa)检验评分者间信度，Kappa > 0.75 表示一致性极佳。

2.4. 统计分析

采用 SPSS 25.0 软件进行数据分析。采用 Shapiro-Wilk 检验对所有计量资料进行正态性分析，符合正态分布的计量资料以均数 ± 标准差表示，组间整体比较采用单因素方差分析(ANOVA)，组间两两比较采用 Tukey HSD 检验；不符合正态分布的计量资料以中位数(四分位数间距)表示，组间比较采用克鲁斯

卡尔–沃利斯 H 检验。$P < 0.05$ 为差异具有统计学意义。

3. 结果

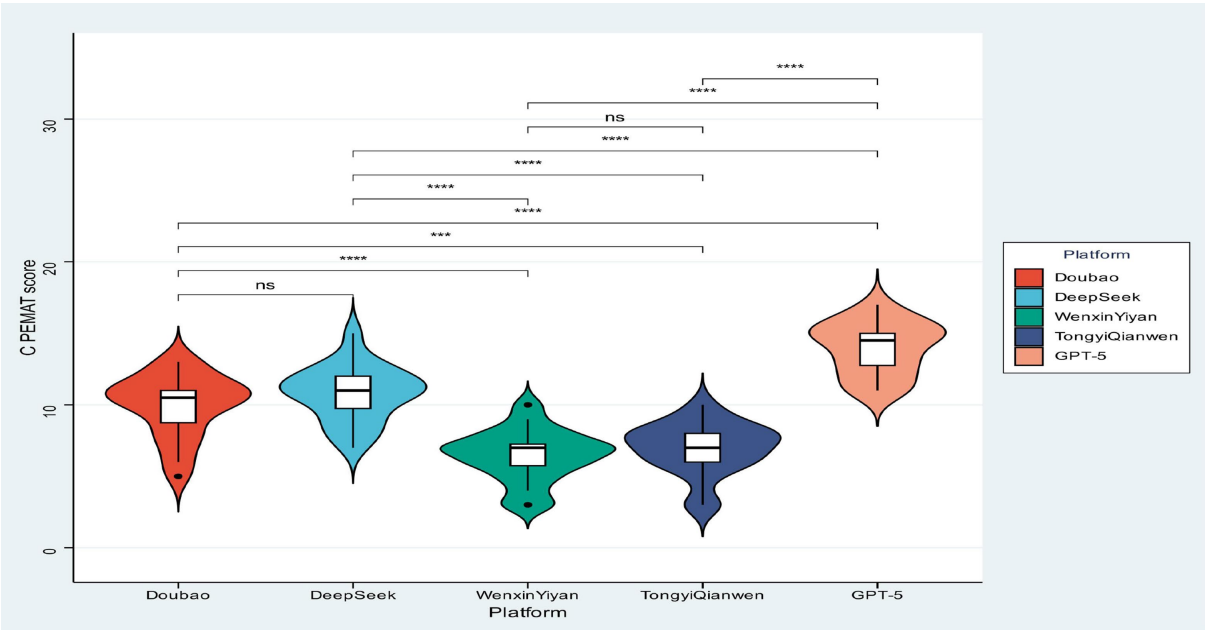
3.1. 样本基线特征

本研究共纳入 100 篇有效文本，每模型 20 篇。5 个核心主题：基础认知与选择、植入手术与术前准备、使用维护与操作规范、并发症与处理、日常护理与随访；各 20 篇，样本分布均衡。

3.2. 不同大型语言模型的表现差异

3.2.1. 质量指标

各模型间 C-PEMAT-P 和 GQS 评分差异均有统计学意义(均 $P < 0.001$)。GPT-5 在两项指标中均获最高评分(C-PEMAT-P: 13.95 ± 1.85; GQS: 4.45 ± 0.51)，在 C-PEMAT-P 评分中，深度求索(10.80 ± 1.94)位列第二，豆包(9.95 ± 2.19)位列第三；在 GQS 评分中，豆包(3.60 ± 0.50)位列第二，深度求索(3.35 ± 0.49)位列第三。文心一言评分最低(C-PEMAT-P: 6.45 ± 1.82; GQS: 1.75 ± 0.64)，显著低于其他模型(事后 Tukey 检验，均 $P < 0.05$) (表 1)。小提琴图进一步显示，GPT-5 和豆包的评分分布集中且整体分值较高，说明其输出质量稳定；而文心一言和通义千问的评分分布离散且整体偏低，输出质量波动较大(图 1、图 2)。



注：C-PEMAT-P (图 1)和 GQS 评分(图 2)所有组间比较均采用 Tukey HSD 事后检验，统计标记含义：ns 代表两组间差异无统计学意义($P \geq 0.05$)；***代表两组间差异有极显著统计学意义($P < 0.001$)；****代表两组间差异有极显著统计学意义($P < 0.0001$)。

Figure 1. Violin plot of C-PEMAT-P scores of infusion port popular science texts generated by different large language models
图 1. 不同大语言模型生成输液港科普文本的 C-PEMAT-P 评分小提琴图

3.2.2. 可读性指标

除 GFOG 外($\chi^2 = 3.51$, $P = 0.477$)，其余可读性指标在各模型间差异均有统计学意义(均 $P < 0.001$)。豆包与 GPT-5 的自动可读性指数(ARI)无显著统计学差异($P > 0.05$)，分别为 15.98 (13.83, 18.70)和 15.93 (15.00, 16.89)，语言复杂度并列最高；文心一言的 ARI 最低[11.70 (11.01, 12.19)]，语言相对简单。FRES

以文心一言最高[49.50 (43.50, 52.00)]，GPT-5 最低[35.50 (32.75, 38.25)]。通义千问的句子长度(CL)最长[14.86 (13.48, 15.41)]，文心一言最短[11.69 (10.91, 12.56)] (表 1)。

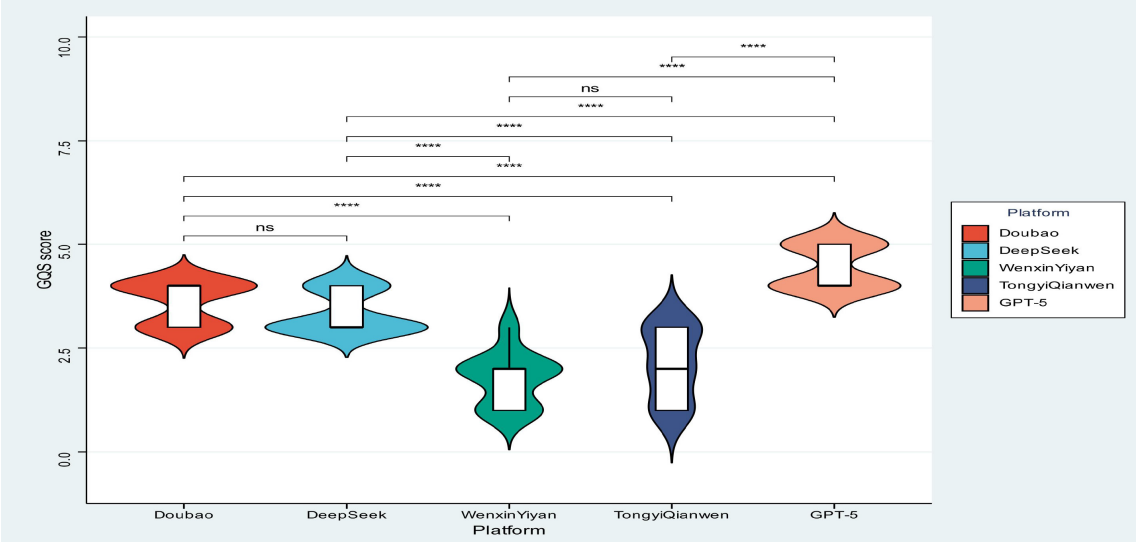


Figure 2. Violin plot of GQS scores of infusion port popular science texts generated by different large language models
图 2. 不同大语言模型生成输液港科普文本的 GQS 评分小提琴图

Table 1. Comparison of quality and readability indicators among different large language models
表 1. 不同大型语言模型质量与可读性指标比较

指标	总计 (n = 100)	深度求索 (n = 20)	豆包 (n = 20)	GPT-5 (n = 20)	通义千问 (n = 20)	文心一言 (n = 20)	统计量	P 值
C-PEMAT-P (均数 ± 标准差)	9.62 ± 3.34	10.80 ± 1.94	9.95 ± 2.19	13.95 ± 1.85	6.95 ± 1.82	6.45 ± 1.82	F = 50.36	<0.001
GQS (均数 ± 标准差)	3.05 ± 1.17	3.35 ± 0.49	3.60 ± 0.50	4.45 ± 0.51	2.10 ± 0.85	1.75 ± 0.64	F = 65.52	<0.001
ARI [M (Q_1, Q_3)]	13.72 (12.05, 16.12)	12.64 (11.55, 13.99)	15.98 (13.83, 18.70)	15.93 (15.00, 16.89)	13.38 (12.85, 14.44)	11.70 (11.01, 12.19)	χ^2 = 38.52#	<0.001
FRES [M (Q_1, Q_3)]	39.00 (34.75, 49.25)	45.50 (38.50, 54.25)	38.50 (30.00, 45.50)	35.50 (32.75, 38.25)	37.00 (32.00, 47.25)	49.50 (43.50, 52.00)	χ^2 = 19.16#	<0.001
GFOG [M (Q_1, Q_3)]	13.45 (12.28, 14.53)	13.05 (12.10, 14.27)	13.75 (12.45, 15.55)	14.05 (13.00, 14.45)	13.60 (12.40, 14.22)	12.85 (11.80, 14.05)	χ^2 = 3.51#	0.477
FKGL [M (Q_1, Q_3)]	12.55 (11.03, 13.85)	12.02 (10.50, 12.89)	13.34 (12.12, 15.73)	14.11 (13.61, 14.72)	12.22 (10.99, 12.64)	10.84 (10.39, 11.59)	χ^2 = 32.88#	<0.001
CL [M (Q_1, Q_3)]	13.43 (11.72, 14.85)	12.00 (11.20, 13.56)	13.64 (12.61, 15.60)	14.24 (13.36, 15.29)	14.86 (13.48, 15.41)	11.69 (10.91, 12.56)	χ^2 = 26.06#	<0.001
SMOG [M (Q_1, Q_3)]	11.57 (10.56, 12.79)	11.37 (10.41, 12.15)	12.25 (11.57, 14.11)	12.67 (12.13, 13.64)	10.84 (9.82, 11.57)	10.57 (10.18, 12.00)	χ^2 = 23.97#	<0.001
LW [M (Q_1, Q_3)]	64.50 (61.00, 68.00)	66.00 (65.00, 69.50)	63.00 (59.50, 65.25)	61.50 (59.75, 63.00)	65.00 (62.50, 69.00)	70.00 (65.75, 74.00)	χ^2 = 28.44#	<0.001

注：F = 方差分析统计量；# = 克鲁斯卡尔 - 沃利斯 H 检验统计量；SD = 标准差；M = 中位数；Q_1 = 第一四分位数；Q_3 = 第三四分位数。

3.3. 不同健康教育主题的表现差异

3.3.1. 质量指标

5 个核心主题间的 C-PEMAT-P 和 GQS 评分均无显著差异。这表明，无论科普主题的复杂程度如何，高质量模型都能保持稳定的内容产出水准(表 2)。

3.3.2. 可读性指标

FRES、GFOG、FKGL、CL 和 SMOG 在各主题间差异有统计学意义(均 $P < 0.05$)，ARI ($\chi^2 = 9.21$, $P = 0.056$)和 LW ($\chi^2 = 4.83$, $P = 0.305$)无显著差异。“日常护理与随访”的 FRES 最高[51.00 (39.75, 55.25)]，可读性最佳；“基础认知与选择”的 ARI [15.93 (13.67, 18.25)]和 FKGL [13.93 (12.70, 15.59)]最高，语言复杂度最高(表 2)。

Table 2. Comparison of quality and readability indicators among different themes
表 2. 不同主题质量与可读性指标比较

指标	总计 (n = 100)	基础认知与选择 (n = 20)	植入手术与术前准备 (n = 20)	使用维护与操作规范 (n = 20)	并发症与处理(n = 20)	日常护理与随访(n = 20)	统计量	P 值
C-PEMAT-P (均数 ± 标准差)	9.62 ± 3.34	10.05 ± 3.27	9.85 ± 3.36	9.35 ± 3.25	9.70 ± 3.13	9.15 ± 3.88	F = 0.23	0.918
GQS (均数 ± 标准差)	3.05 ± 1.17	3.00 ± 1.30	3.05 ± 0.89	3.10 ± 1.33	3.10 ± 1.25	3.00 ± 1.12	F = 0.04	0.998
ARI [M (Q_1, Q_3)]	13.72 (12.05, 16.12)	15.93 (13.67, 18.25)	13.28 (12.61, 14.59)	14.80 (11.59, 15.90)	14.25 (12.18, 16.20)	12.75 (11.36, 14.08)	$\chi^2 = 9.21$#	0.056
FRES [M (Q_1, Q_3)]	39.00 (34.75, 49.25)	35.50 (27.50, 39.50)	40.00 (36.00, 48.00)	44.50 (36.00, 50.25)	37.00 (31.75, 42.50)	51.00 (39.75, 55.25)	$\chi^2 = 15.68$#	0.003
GFOG [M (Q_1, Q_3)]	13.45 (12.28, 14.53)	14.20 (13.57, 15.50)	13.65 (13.02, 14.55)	13.05 (12.10, 14.35)	13.15 (12.50, 14.20)	12.50 (11.20, 13.07)	$\chi^2 = 15.28$#	0.004
FKGL [M (Q_1, Q_3)]	12.55 (11.03, 13.85)	13.93 (12.70, 15.59)	12.40 (11.24, 13.38)	12.73 (10.86, 13.76)	12.85 (11.21, 13.60)	11.29 (10.75, 12.89)	$\chi^2 = 11.53$#	0.021
CL [M (Q_1, Q_3)]	13.43 (11.72, 14.85)	14.24 (12.44, 15.60)	13.44 (11.99, 13.90)	13.04 (11.66, 15.29)	14.34 (13.21, 15.35)	11.51 (10.93, 13.30)	$\chi^2 = 14.62$#	0.006
SMOG [M (Q_1, Q_3)]	11.57 (10.56, 12.79)	12.54 (11.57, 14.11)	11.57 (11.07, 12.71)	11.67 (10.32, 12.91)	11.37 (9.75, 12.41)	10.69 (9.39, 12.16)	$\chi^2 = 13.00$#	0.011
LW [M (Q_1, Q_3)]	64.50 (61.00, 68.00)	63.00 (58.00, 65.00)	64.50 (62.00, 66.25)	65.50 (61.75, 68.00)	65.00 (62.75, 71.00)	64.00 (61.25, 72.25)	$\chi^2 = 4.83$#	0.305

注：F = 方差分析统计量；# = 克鲁斯卡尔 - 沃利斯 H 检验统计量；SD = 标准差；M = 中位数；Q_1 = 第一四分位数；Q_3 = 第三四分位数。

3.4. 质量与可读性的相关性

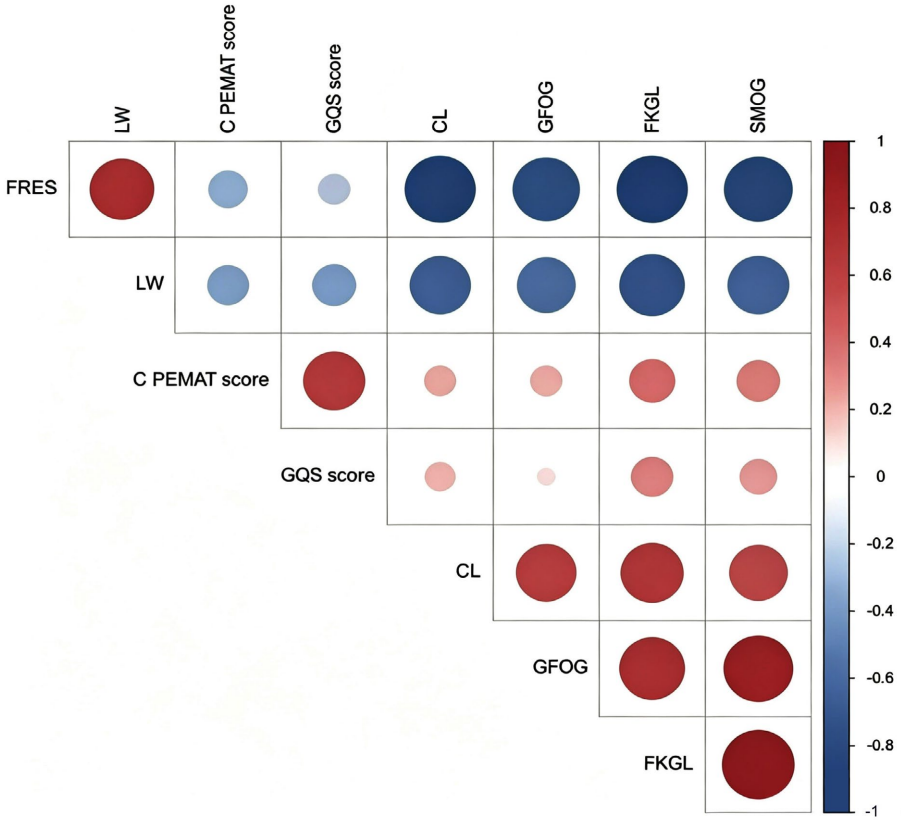
斯皮尔曼相关性分析显示，两项质量指标之间呈中等正相关(GQS 与 C-PEMAT-P：$r = 0.63$，$P < 0.001$)，说明不同评估工具对文本质量的判断具有一致性。可读性指标中，ARI 与 FKGL 呈强正相关($r = 0.85$, $P < 0.001$)，FRES 与 GFOG 呈强负相关($r = -0.82$, $P < 0.001$)。质量指标与可读性指标相关性较弱：C-PEMAT-P 与 SMOG 呈弱负相关($r = -0.18$, $P = 0.042$)，GQS 与 LW 呈弱正相关($r = 0.21$, $P = 0.028$)。这进一步证实，文本的专业质量与可读性是两个相对独立的评价维度(图 3)。

4. 讨论

4.1. 大型语言模型的表现差异及潜在机制

本研究发现，主流大型语言模型生成的植入式静脉输液港(IVAPs)科普文本质量存在显著差异，其中 GPT-5 综合表现最优，深度求索在内容可理解性与可操作性维度表现更优，豆包在整体内容流畅度与结构完整性维度表现更优，二者综合表现显著优于文心一言和通义千问。这种差异不仅反映了各模型在底层算法架构上的区别，更凸显了训练数据集在医疗垂直领域的广度与深度对输出质量的决定性作用[14]。GPT-5 的优异表现可能与其预训练语料库中深度融合了海量高质量医疗指南、循证医学文献及标准化患者教育材料有关，使其能够跨越简单的常识拼凑，生成具备高度结构化、准确性且具有临床可操作性的

专业内容。例如，在解释输液港冲管这一关键临床操作时，GPT-5 能够精准给出“使用≥10 mL 生理盐水脉冲式冲管后正压封管”的具体循证指导。相比之下，文心一言等模型仅泛泛提及“规范冲管”，缺乏具体的操作细节和剂量规范，这直接导致其在衡量材料可操作性的 C-PEMAT-P 评分中表现不佳。



注：本图展示 8 项核心评估指标间的斯皮尔曼相关性分析结果。横、纵坐标均为纳入分析的评估指标，包含 2 项质量指标(C-PEMAT score, GQS score)和 6 项可读性指标(FRES, LW, CL, GFOG, FKGL, SMOG)。右侧颜色条代表斯皮尔曼相关系数(r)的取值范围(−1~1)，红色系表示正相关，蓝色系表示负相关，颜色深浅与相关性强弱正相关；圆圈大小与相关系数的绝对值成正比，圆圈越大代表两项指标间的相关性越强。

Figure 3. Spearman correlation heatmap of quality and readability indicators of infusion port popular science texts
图 3. 输液港科普文本质量与可读性指标的斯皮尔曼相关性热图

此外，本研究揭示了一个关键的现象：文本的“可读性”与“质量”并不完全对等。文心一言虽然在 FRES 指标上得分最高，但其专业质量评分却垫底。这一现象警示我们：单纯为了追求通俗易懂而过度简化内容，极易导致核心医疗信息的遗漏或失真，最终损害科普材料的准确性与完整性，进而损害科普材料的准确性和完整性[15]。这一发现有力地挑战了传统健康沟通中“可读性越高即质量越好”的刻板印象，强调在高度专业化的医疗科普中，必须在“语言的简洁性”与“医学的专业性”之间寻求科学的动态平衡[16]。

4.2. 临床意义与优化策略

研究表明，健康教育主题的复杂程度会显著影响文本的可读性，但并不会削弱高质量模型的信息产出质量。对于临床医护人员而言，这意味着可以信任并放心地将经过筛选的高质量模型(如 GPT-5)作为数字健康干预的辅助引擎。本研究结论仅适用于中文通用大语言模型生成的植入式静脉输液港主题科普文

本，不可直接推广至其他语言版本、医疗垂直领域专用模型或其他疾病/医疗装置的科普内容。在实际临床应用中，直接将模型生成的原始文本发送给患者并非最佳实践。研究指出，针对患者自身健康素养量身定制的数字干预策略，能够更有效地提升患者对关键医疗信息的吸收和转化效率[17]。

因此，建议在植入式静脉输液港科普实践中采用差异化定向健康传播思路，依托数字化宣教载体优化健康信息供给。循证研究表明，相较于传统常规教育模式，多媒体化、个性化的健康宣教可有效提升长期静脉通路装置患者的知识掌握水平与自我管理能力，减少导管相关并发症，是优化护患健康沟通的有效路径[18]。

4.3. 质量与可读性的相对独立性及其实践启示

在人工智能技术与临床护理实践深度融合的时代背景下，模型类型依然是决定静脉输液港科普文本质量的核心基石。GPT-5、豆包和深度求索凭借其强大的逻辑推理和信息整合能力，成为现阶段临床科普的优选工具。更为重要的是，医护人员必须深刻认识到“文本质量”与“可读性”是两个相对独立的评价维度——高质量的内容未必易于理解，而极易读的内容可能缺乏指导深度。未来在优化健康教育材料时，应建立“质量保底、可读性动态调节”的新范式，将客观的语言学指标与患者真实报告结局(PROs)紧密结合，重视倾听患者主观体验与真实健康需求[19]，以此构建更具温度和精度的循证健康教育体系。

4.4. 优势与局限性

本研究具有多重优势：第一：系统且全面地评估了当前主流的大语言模型；第二：精准聚焦于“植入式静脉输液港”这一临床需求高但既往 AI 研究较少涉足的专业细分领域；第三：采用了经过国际验证、具有高信效度的可读性和质量双重客观评估体系，研究结果具有较强的科学性；第四，本研究仅评估了中文科普文本，未涉及多语言版本的生成质量与可读性差异；第五，研究主题局限于植入式静脉输液港，未拓展至其他需要长期自我管理的医疗装置或慢性疾病领域。

本研究亦存在一定的局限性：第一，受限于研究开展时间，仅纳入了 5 款现阶段的主流通用模型。随着 AI 技术的快速迭代，未来仍需持续纳入新兴的或医疗垂直领域专用的语言模型，以扩大结论的适用范围；第二，本评估框架主要依赖于客观语言学参数和专家的外部质量评分，未能纳入患者报告结局及最终的临床终点事件；第三，未充分考量患者群体的个体化差异，使得统一的可读性评估在面对高度异质性的临床真实世界时，普适性受到一定限制。

4.5. 未来研究方向

基于上述局限，未来的研究可从以下三个维度进行深化：第一，横向拓展模型应用与主题范围，将评估对象延伸至其他需要高度精细化自我管理的专业医疗装置；第二，纵向构建多维度评估体系，将客观文本指标、专家的外部质量评分与患者的主观认知反馈有机整合，通过认知访谈、可用性测试、焦点小组讨论等定性研究方法，系统评估患者对不同 LLM 生成内容的真实理解程度、接受度和行为改变意愿，形成包含“客观指标 + 专家评估 + 患者反馈”的三角验证评估体系；同时纳入患者报告结局(PROs)及导管相关并发症发生率等长期临床护理结局，实现真正的闭环验证；第三，探索结合检索增强生成(RAG)技术，设计嵌入临床信息系统的个性化科普生成工具，根据患者具体特征动态调整文本可读性，甚至拓展语音、视频等多模态输出形式，真正实现个性化的精准健康沟通。

5. 结论

综上所述，大语言模型的底层架构和训练数据质量是决定植入式静脉输液港科普文本专业质量的核心要素。其中，GPT-5 在医学精准度与指导框架的结构化方面表现最优，深度求索在内容可操作性维度

表现突出，豆包在内容流畅度与结构完整性维度表现优异，三者可作为现阶段临床健康教育的最优辅助工具。研究进一步揭示，科普主题的复杂度显著影响文本的可读性，但并不削弱高质量模型的内容产出水准。具体而言，侧重实操的“日常护理与随访”模块文本普适性最高，而涉及专业机制的“基础认知与选择”模块则呈现出较高的阅读门槛。本研究结论仅适用于中文通用大语言模型生成的植入式静脉输液港主题科普内容，在推广至其他场景时需谨慎验证。

鉴于质量与可读性是衡量科普材料的两个独立维度，两者的协同优化对于提升健康沟通效能至关重要。在未来的临床护理实践中，医护人员应积极践行“人机协同”理念：优先调用优效模型搭建专业知识骨架，再紧密结合科普主题的复杂度与特定患者群体的健康素养，进行个性化的可读性降维与语义润色。这一转化策略不仅能深化患者对专业知识的实质性吸收，更能有效赋能其居家自我管理能力。

展望未来，本领域需开展多中心、前瞻性的队列研究，将真实世界中的患者报告结局(PROs)与客观临床指标纳入评价体系，同时引入定性研究方法收集患者的真实使用体验与认知反馈，从而进一步确证并挖掘大型语言模型在重塑现代医疗健康教育生态中的核心应用价值。

声　明

根据《赫尔辛基宣言》及相关学术伦理指南，无需获得伦理审批。

利益冲突

作者声明无潜在利益冲突。

参考文献

[1] 刘鹏, 吴巍巍. 静脉输液港植入与管理多学科专家共识(2023 版) [J]. 中国普通外科杂志, 2023, 32(6): 799-814.

[2] 何越, 孙艳萍, 李宁, 沈继龙. 血液恶性肿瘤患者应用 PICC 与植入式静脉输液港的效果比较[J]. 中华护理杂志, 2012, 47(11): 1001-1003.

[3] 王建新, 唐甜甜, 谢艳丽. 植入式静脉输液港常见并发症的临床分析[J]. 护士进修杂志, 2012, 27(10): 958-960.

[4] Paterick, T.E., Patel, N., Tajik, A.J. and Chandrasekaran, K. (2017) Improving Health Outcomes through Patient Education and Partnerships with Patients. *Baylor University Medical Center Proceedings*, **30**, 112-113. https://doi.org/10.1080/08998280.2017.11929552

[5] Daraz, L., Morrow, A.S., Ponce, O.J., Farah, W., Katabi, A., Majzoub, A., *et al*. (2018) Readability of Online Health Information: A Meta-Narrative Systematic Review. *American Journal of Medical Quality*, **33**, 487-492. https://doi.org/10.1177/1062860617751639

[6] Radford, A., Narasimhan, K., Salimans, T., *et al*. (2018) Improving Language Understanding by Generative Pre-Training. OpenAI. (Preprint)

[7] OpenAI (2023) GPT-4 Technical Report. arXiv: 2303.08774. https://arxiv.org/abs/2303.08774

[8] Liu, S., McCoy, A.B. and Wright, A. (2025) Improving Large Language Model Applications in Biomedicine with Retrieval-Augmented Generation: A Systematic Review, Meta-Analysis, and Clinical Development Guidelines. *Journal of the American Medical Informatics Association*, **32**, 605-615. https://doi.org/10.1093/jamia/ocaf008

[9] Parameswaran, V., Bernard, J., Bernard, A., Deo, N., Tsung, S., Lyytinen, K., *et al*. (2025) Evaluating Large Language Models and Retrieval-Augmented Generation Enhancement for Delivering Guideline-Adherent Nutrition Information for Cardiovascular Disease Prevention: Cross-Sectional Study. *Journal of Medical Internet Research*, **27**, e78625. https://doi.org/10.2196/78625

[10] DuBay, W.H. (2004) The Principles of Readability (ED490073). ERIC.

[11] Perni, S., Rooney, M.K., Horowitz, D.P., Golden, D.W., McCall, A.R., Einstein, A.J., *et al*. (2019) Assessment of Use, Specificity, and Readability of Written Clinical Informed Consent Forms for Patients with Cancer Undergoing Radiotherapy. *JAMA Oncology*, **5**, e190260. https://doi.org/10.1001/jamaoncol.2019.0260

[12] Shoemaker, S.J., Wolf, M.S. and Brach, C. (2014) Development of the Patient Education Materials Assessment Tool (PEMAT): A New Measure of Understandability and Actionability for Print and Audiovisual Patient Information. *Patient Education and Counseling*, **96**, 395-403. https://doi.org/10.1016/j.pec.2014.05.027

[13] Bernard, A., Langille, M., Hughes, S., Rose, C., Leddin, D. and Veldhuyzen van Zanten, S. (2007) A Systematic Review of Patient Inflammatory Bowel Disease Information Resources on the World Wide Web. *The American Journal of Gastroenterology*, **102**, 2070-2077. https://doi.org/10.1111/j.1572-0241.2007.01325.x

[14] 肖仰华, 徐一丹. 大规模生成式语言模型在医疗领域的应用: 机遇与挑战[J]. 医学信息学杂志, 2023, 44(9): 1-11.

[15] 王蕾, 汪秋伊, 李星, 等. 网络健康信息可读性评估研究现状及展望[J]. 医学信息学杂志, 2020, 41(12): 20-25, 40.

[16] Lee, H., Kim, S., Kim, S., Seo, J., Kim, W.H., Kim, J., *et al*. (2025) Readability versus Accuracy in LLM-Transformed Radiology Reports: Stakeholder Preferences across Reading Grade Levels. *La Radiologia Medica*, **130**, 1986-1999. https://doi.org/10.1007/s11547-025-02098-5

[17] Hovingh, J.W., Elderson-van Duin, C., Kuipers, D.A., van Rood, Y., Ludden, G.D.S., Hanssen, D.J.C., *et al*. (2025) Tailoring for Health Literacy in the Design and Development of Ehealth Interventions: Systematic Review. *JMIR Human Factors*, **12**, e76172. https://doi.org/10.2196/76172

[18] Basso, I., El Motarajji, S., Ferrari, M., Airoldi, C., Durante, A., Brovarone, S., *et al*. (2026) The Effectiveness of a Multimedia Education versus a Standard Education Program in the Self-Management of Central Venous Catheters for Long-Term Use: A Systematic Review. *The Journal of Vascular Access*, **27**, 885-894. https://doi.org/10.1177/11297298251378618

[19] 蒋璐璐, 王喜益, 徐洁慧, 等. 智能交互式护理信息支持系统的构建及在乳腺癌患者中的应用研究[J]. 中华护理杂志, 2023, 58(6): 654-661.

附 录

Table S1. List of popular science related questions about implantable venous access ports
表 S1. 植入式静脉输液港科普相关问题清单

主题(Theme)	序号(No.)	具体问题(Specific Question)
基础认知与选择	1	输液港是干什么用的？
	2	会不会留下明显疤痕或鼓包？
	3	和 PICC/中心导管相比哪个更舒适省事？
	4	我适不适合放输液港，治疗结束后怎么办？
置入手术与准备	1	置入手术疼吗？用什么麻醉？
	2	手术一般多久，当天能回家吗？
	3	术前需要空腹、停药或做哪些检查？
	4	输液港放在哪个位置，我能参与选择吗？
使用维护与规范	1	不用治疗时，多久维护一次比较合适？
	2	每次穿刺会疼吗？可以用表麻贴减痛吗？
	3	能在社区或外院维护吗？需要携带什么材料？
	4	输液港能抽血、打造影剂或做高压注射吗？
并发风险与处理	1	出现红肿、渗液或发热时我该怎么办？
	2	回抽不出血但能输液，需要担心吗？
	3	我能做些什么来降低感染和血栓风险
	4	外渗、移位或断裂有什么警示信号？
生活管理与随访	1	洗澡、游泳、泡温泉可以吗？如何防护？
	2	运动、提重物、背包或系安全带会影响吗？
	3	过安检、坐飞机，做 X 线/CT/MRI 行不行？
	4	何时考虑拔除？随访频率与预约流程是怎样的？