DOI: 10.16781/j.CN31-2187/R.20250542

·论　著·

大语言模型在围手术期肺保护通气策略建议中的比较研究

吴　奇，陆　军，张焱琳，王晓琳，杨　涛，卞金俊，薄禄龙*
海军军医大学（第二军医大学）第一附属医院麻醉科，上海 200433

［摘要］ **目的**　评估不同大语言模型（LLM）在围手术期肺保护通气策略建议中的专业适配性。**方法**　采用横断面研究方法，以 2019 年发布的《手术患者肺保护性通气策略的国际专家组推荐规范》为标准，选取 8 种 LLM 并通过统一提示词生成围手术期肺保护通气策略建议。基于总偏离评分（TDS）、医疗人工智能质量评估工具（QAMAI），采取 8 种 LLM 和麻醉科医师评价的方式进行结果校正。**结果**　LLM 生成建议与指南一致性较强，但均未达到内容完全吻合。LLM 评价结果中 ChatGPT o1 表现最佳，TDS 评分为 1.5（1.0，2.0）分，QAMAI 评分为 25.5（22.2，27.8）分；Gemini 2.0 pro 表现最差，TDS 评分为 3.0（2.0，5.0）分，QAMAI 评分为 21.0（16.8，23.8）分。在国内 LLM 中，DeepSeek R1 表现最佳，TDS 评分为 2.0（1.5，2.8）分，QAMAI 评分为 25.5（19.2，28.2）分。麻醉科医师同行评议结果与 LLM 评价结果一致。**结论**　LLM 在围手术期肺保护通气策略建议中呈现异质性，ChatGPT o1 与 DeepSeek R1 在围手术期肺保护通气策略建议中表现出较高的准确性，具有潜在的临床参考价值。

［关键词］ 大语言模型；肺保护通气；围手术期管理；麻醉安全；人工智能
［引用本文］ 吴奇，陆军，张焱琳，等. 大语言模型在围手术期肺保护通气策略建议中的比较研究［J］. 海军军医大学学报，2025，46（12）：1582-1589. DOI: 10.16781/j.CN31-2187/R.20250542.

Comparison of large language models for perioperative lung-protective ventilation recommendations

WU Qi, LU Jun, ZHANG Yanlin, WANG Xiaolin, YANG Tao, BIAN Jinjun, BO Lulong*
Department of Anesthesiology, The First Affiliated Hospital of Naval Medical University (Second Military Medical University), Shanghai 200433, China

［**Abstract**］ **Objective**　To evaluate the professional adaptability of different large language models (LLMs) in generating recommendations for perioperative lung-protective ventilation. **Methods**　A cross-sectional study was conducted using *Lung-protective ventilation for the surgical patient: international expert panel-based consensus recommendations* published in 2019 as the standard. Eight LLMs were selected to generate ventilation strategy recommendations via standardized prompts. Outcomes were calibrated through assessment by the 8 LLMs and anesthesiologists based on total disagreement score (TDS) and quality assessment of medical artificial intelligence (QAMAI). **Results**　Recommendations generated by LLMs showed strong alignment with the expert consensus but failed to achieve full concordance across all items. Assessment results by LLMs indicated that ChatGPT o1 performed best (TDS 1.5［1.0, 2.0］, QAMAI score 25.5［22.2, 27.8］), while Gemini 2.0 pro showed the poorest performance (TDS 3.0［2.0, 5.0］, QAMAI score 21.0［16.8, 23.8］). Among domestic LLMs, DeepSeek R1 achieved the best scores (TDS 2.0［1.5, 2.8］, QAMAI score 25.5［19.2, 28.2］). Peer review results by anesthesiologists were consistent with LLM assessment findings. **Conclusion**　The recommendations of LLMs for perioperative lung-protective ventilation strategies exhibit heterogeneity, with ChatGPT o1 and DeepSeek R1 demonstrating relatively higher accuracy and potential clinical utility.

［**Key words**］ large language model; lung-protective ventilation; perioperative management; anesthesia safety; artificial intelligence
［**Citation**］ WU Q, LU J, ZHANG Y, et al. Comparison of large language models for perioperative lung-protective ventilation recommendations［J］. Acad J Naval Med Univ, 2025, 46(12): 1582-1589. DOI: 10.16781/j.CN31-2187/R.20250542.

肺保护通气策略是围手术期呼吸管理的核心干预措施[1]。小潮气量通气、驱动压控制、呼气末正压和肺复张手法等协同使用可使术后肺部并发症风险降低 37%～52%[2]。然而，麻醉临床实践中仍存在指南依从性不足等特点。近年来，大语言模型（large language model，LLM）通过学习海量医

［收稿日期］ 2025-08-12　［接受日期］ 2025-10-28
［作者简介］ 吴　奇，硕士生，住院医师. E-mail: 19951152133@163.com
*通信作者（Corresponding author）. E-mail: bartbo@smmu.edu.cn

学文献与构建知识图谱，在临床决策支持领域展现出巨大潜力。但是，不同 LLM 在算法架构、训练数据时效性及医学逻辑推理能力等方面存在显著异质性，其在专业领域建议的可靠性仍不明确[3]。既往相关研究大多聚焦于 LLM 对通用医学知识的评估，鲜有研究系统探讨其在麻醉学领域的应用与实践[4-8]。Saxena 等[9]对国外几种主流 LLM 在生成围手术期神经功能认知障碍建议方面的表现进行了评估。结果表明，不同模型生成的建议与公布的指南结果不完全一致且缺乏引用来源。本研究拟对比国内外 8 种 LLM 在围手术期肺保护通气策略建议方面的表现，评估其准确性和优劣，为麻醉医师选择可靠的 LLM 提供参考依据。

1　资料和方法

1.1　*研究设计与数据收集*　本研究为基于互联网的横断面分析研究。选择使用较为广泛的 8 种 LLM，为避免版本更新对结果产生影响，所有 LLM 于 2025 年 2 月 23－24 日进行访问并检索。其中，文心大模型 4.0、通义千问、Kimi、豆包、新青年麻醉 AI 助手（高级版）均在其对应的官方网站进行检索，ChatGPT o1、Gemini 2.0 pro、DeepSeek R1 通过微软 Edge 浏览器 sider 插件进行检索。所有 LLM 版本均为 2025 年 2 月 23－24 日最新版本。部分 LLM 存在版本更新，本研究不纳入后续新版本。

由于纳入国外 LLM，因此统一向所有 LLM 输入提示词：Please create a detailed and precise table that outlines a comprehensive bundle of strategies for lung-protective ventilation specifically tailored for surgical patients。该提示词由 3 位资深麻醉科副主任医师共同商议决定。为避免反复提问对 LLM 结果产生影响，每一项对话请求均在全新对话框中完成，且只选择各模型回复的第一个答案。最终，共生成 8 个 LLM 输出建议的表格，在删除标识该模型名称的信息后将其分别命名为模型 1～8，然后合并至一个 Word 文档。

1.2　*评价方法与评分标准*　将该 Word 文档上传至 8 种 LLM 进行评价。评价方法如下：以 2019 年发布的《手术患者肺保护性通气策略的国际专家组推荐规范》（以下简称指南）为参照标准，其包括术前风险评估、术中通气策略、肺复张策略、麻醉苏醒等方面的 22 项建议和 4 项声明[10-11]。LLM 在学习该参考标准后，根据总偏离评分（total disagreement score，TDS）、医疗人工智能质量评估工具（quality assessment of medical artificial intelligence，QAMAI）的评分标准，对模型 1～8 进行评分并给出评分说明[12]。此外，按照上述标准，由 2 位麻醉科副主任医师对模型 1～8 进行 TDS 及 QAMAI 评分，如果评分存在明显争议（TDS 差异≥2 分，QAMAI 评分差异≥5 分）则纳入第 3 位麻醉科副主任医师的评分[13]。

TDS 将 LLM 的回答与指南的偏离程度划分为 0 分（完全一致）～3 分（完全偏离）。评分标准：0 分为无异议，1 分为轻微异议（答案缺少非关键性细节），2 分为中等程度异议（存在 1 个或多个答案细节错误，但这些错误对患者的治疗结果无关键性影响），3 分为重大异议（答案缺少或错误提供了可能对患者治疗结果至关重要的信息）。TDS 从术前风险评估、术中通气策略及其他策略 3 个方面分别进行评分，每个 LLM 的最终总分为 0～9 分。

QAMAI 评分从准确性、清晰度、相关性、完整性、提供来源和有用性共 6 个方面反映 LLM 与指南的一致程度，每个方面得分为 1 分（完全不同意）～5 分（完全同意）。评分标准：QAMAI 评分 6～11 分为质量差，表示该模型提供的信息大多不可靠或不完整，需要立即改进；12～17 分为质量一般，表示该模型提供了一些有用信息，但仍有大量改进空间；18～23 分为质量良好，表示该模型提供的信息大多可靠且完整，但可能仍有部分方面有待完善；24～29 分为质量非常好，表示该模型在大多数方面提供了可靠且完整的信息，仅有少数方面需要改进；30 分为质量优秀，表示该模型提供了高度可靠且完整的信息[9]。

最后，根据 TDS 和 QAMAI 评分对各模型生成的建议进行综合评级，分为完全符合、大致符合和不符合 3 类：TDS 总评分为 0 分、QAMAI 总评分为 24～30 分，属于完全符合；TDS 总评分为 1～3 分、QAMAI 总评分为 18～23 分，属于大致符合；TDS 总评分≥4 分，QAMAI 总评分为 6～17 分，属于不符合。统计各 LLM 中不符合的个数，对不符合个数 / 总个数比值进行排序，获得 LLM 评价及麻醉科医师同行评价对各 LLM 的结论认可度排序。

1.3　统计学处理　应用SPSS 27.0软件和GraphPad Prism 9软件进行统计学分析。计量资料以$M(Q_1, Q_3)$表示，组间比较采用Kruskal-Wallis H检验，检验水准（α）为0.05。使用Cliff's δ效应量描述不同LLM评分差异的实际意义[14-15]。

2　结　果

2.1　LLM生成的围手术期肺保护通气策略建议及其与指南的对比　8种LLM生成的围手术期肺保护通气策略建议及指南中相应描述见表1。

表1　大语言模型生成围手术期肺保护通气策略建议及指南推荐建议

Tab 1　Recommendations for perioperative lung-protective ventilation by large language models and by guidelines

Item	Ernie Bot4.0	Doubao	Kimi	Tongyi Qianwen	New Youth Anesthesia AI Assistant (advanced version)
Preoperative evaluation protocol and ventilator setting					
Preoperative risk assessment: utilize specialized scoring systems to evaluate the risk of postoperative pulmonary complications	/	/	/	/	++
Initial ventilation setting: tidal volume	6-8 mL•kg^{-1}	6-8 mL•kg^{-1}	6-8 mL•kg^{-1}	6-8 mL•kg^{-1}	6-8 mL•kg^{-1}
PEEP setting	+	5 cmH_2O	5-10 cmH_2O	5-15 cmH_2O	5-8 cmH_2O
Driving pressure setting	/	/	/	+	＜15 cmH_2O
Plateau pressure setting	/	30 cmH_2O	/	＜30 cmH_2O	＜30 cmH_2O
Individualized ventilation setting	+	/	++	++	++
Intraoperative ventilation management protocol					
Position management during anesthesia	/	/	/	+	/
Management during anesthesia induction period: use IPPV/CPAP before induction	/	/	/	/	/
Oxygenation target	SpO_2 94%-98%, with the lowest FiO_2 possible	SpO_2 92%-96%, FiO_2 0.4-0.6	/	SpO_2 88%-92%, with the lowest FiO_2 possible	SpO_2 92%-96%, avoiding FiO_2＞0.6
Management during anesthesia induction period: avoid manual bag reinflation	/	/	/	/	/
Lung recruitment pressure	30-40 cmH_2O, sustained for 30-60 s	30-40 cmH_2O, sustained for 20-40 s	20-30 cmH_2O, sustained for 15-20 s	/	30-40 cmH_2O, sustained for 30-40 s
Hemodynamic monitoring to adjust ventilation and lung recruitment strategy	++	/	/	++	++
Lung recruitment maneuvers: multiple recruitments	++	/	++	++	/
Individualized lung recruitment strategy	++	/	++	++	/
Intraoperative monitoring of dynamic compliance, driving pressure, plateau pressure, and FiO_2	++	/	+	+	/
Ventilator mode	VCV/PCV	VCV/PCV	/	/	/
Inspiratory/expiratory ratio	1∶2-1∶3	1∶2-1∶3	/	/	/
Respiratory rate	+	/	/	/	/
Permissive hypercapnia	/	/	++	/	++
One-lung ventilation setting	/	/	++	/	/
Fluid management	/	/	/	/	/
Postoperative ventilation management protocol and additional consideration					
Management during the recovery period: positioning and maintaining FiO_2＜0.4 during recovery, and providing oxygen therapy as needed after extubation	++	/	+	+	++
Post-extubation care	+	/	+	+	+
Humidification	++	/	/	/	/
Multidisciplinary collaboration	/	/	/	/	++
Sedation and analgesia	/	/	/	/	/

表 1（续）

Item	ChatGPT o1	Gemini 2.0 pro	DeepSeek R1	Guideline[10]
Preoperative evaluation protocol and ventilator setting				
Preoperative risk assessment: utilize specialized scoring systems to evaluate the risk of postoperative pulmonary complications	＋	/	＋	＋＋
Initial ventilation setting: tidal volume	6-8 mL•kg^{-1}	4-8 mL•kg^{-1}	6-8 mL•kg^{-1}	6-8 mL•kg^{-1}
PEEP setting	5 cmH_2O	5-10 cmH_2O	5-10 cmH_2O	5 cmH_2O
Driving pressure setting	/	＜15 cmH_2O	/	≤15 cmH_2O
Plateau pressure setting	＜30 cmH_2O	＜30 cmH_2O	＜30 cmH_2O	＋
Individualized ventilation setting	＋＋	＋＋	＋＋	＋＋
Intraoperative ventilation management protocol				
Position management during anesthesia	＋＋	＋	＋＋	Head-up angle≥30° (beach chair position)
Management during anesthesia induction period: use IPPV/CPAP before induction	/	/	/	＋＋
Oxygenation target	SpO_2 92%-98%, with the lowest FiO_2 possible	SpO_2 92%-96%, avoiding FiO_2＞0.6	SpO_2 92%-96%, FiO_2≤0.8	SpO_2≥94% (FiO_2≤0.4)
Management during anesthesia induction period: avoid manual bag reinflation	/	/	/	＋＋
Lung recruitment pressure	30-40 cmH_2O, sustained for 30 s	/	30-40 cmH_2O, sustained for 30-40 s	30-40 cmH_2O for non-obese patients, 40-50 cmH_2O for obese patients
Hemodynamic monitoring to adjust ventilation and lung recruitment strategy	＋＋	＋＋	＋＋	＋＋
Lung recruitment maneuvers: multiple recruitments	＋＋	/	/	＋＋
Individualized lung recruitment strategy	/	＋＋	/	＋＋
Intraoperative monitoring of dynamic compliance, driving pressure, plateau pressure, and FiO_2	＋＋	＋＋	＋＋	＋＋
Ventilator mode	/	VCV/PCV	PCV	VCV/PCV
Inspiratory/expiratory ratio	/	＋	1:1.5-1:2.5	/
Respiratory rate	＋	＋＋	＋＋	/
Permissive hypercapnia	/	＋＋	/	/
One-lung ventilation setting	/	/	＋	/
Fluid management	/	＋＋	/	/
Postoperative ventilation management protocol and additional consideration				
Management during the recovery period: positioning and maintaining FiO_2＜0.4 during recovery, and providing oxygen therapy as needed after extubation	＋＋	/	＋＋	＋＋
Post-extubation care	/	＋	/	＋
Humidification	/	/	/	/
Multidisciplinary collaboration	＋	/	/	/
Sedation and analgesia	＋＋	＋	/	/

1 cmH_2O＝0.098 kPa. AI: Artificial intelligence; PEEP: Positive end-expiratory pressure; IPPV: Intermittent positive pressure ventilation; CPAP: Continuous positive airway pressure; SpO_2: Peripheral capillary oxygen saturation; FiO_2: Fraction of inspired oxygen; VCV: Volume-controlled ventilation; PCV: Pressure-controlled ventilation. "/": Not mentioned; "＋": Partially mentioned; "＋＋": Fully mentioned.

在术前风险评估与呼吸机设置方面，仅新青年麻醉 AI 助手（高级版）明确建议使用专用评分系统评估术后肺部并发症风险，这与指南内容一致，ChatGPT o1 和 DeepSeek R1 也部分提及该建

议，其余模型均未提及。大多数模型提到呼气末正压通气设置，但具体数值范围有所不同，如 5～15 cmH_2O（1 cmH_2O＝0.098 kPa）（通义千问）、5～8 cmH_2O［新青年麻醉 AI 助手（高级版）］等。通义千问、新青年麻醉 AI 助手（高级版）、ChatGPT o1、Gemini 2.0 pro、DeepSeek R1 等模型提到了驱动压和/或平台压的设置，如驱动压＜15 cmH_2O 或平台压＜30 cmH_2O。另外，多数模型强调了个体化通气设置的重要性，表明需要根据患者的具体情况来调整通气参数。

在术中通气策略方面，所有模型均提及潮气量的设置范围，但在具体数值范围的表述上略有差异，如 Gemini 2.0 pro 表述为 4～8 mL/kg，其他模型均表述为 6～8 mL/kg。多数模型强调了术中监测动态顺应性、驱动压、平台压和 FiO_2 的重要性，并提到了呼吸机通气模式的选择等。各模型在氧合目标和肺复张策略上存在显著差异，如 SpO_2 和 FiO_2 的具体数值范围、肺复张压力的设置等。大多数模型提到血流动力学监测在调整通气及肺复张策略中的重要性。

在术后通气管理方面，除豆包、Gemini 2.0 pro 外，其他模型均提及苏醒期管理的重要性，包括体位、苏醒期吸入氧浓度控制以及拔管后的按需氧疗和护理。

虽然各模型在某些具体策略上存在差异，但整体强调个体化通气设置、术中监测以及苏醒期管理的重要性。部分模型还提出了特定策略或通气参数设置。例如，文心大模型 4.0 提出了湿化防止气道干燥，Kimi、新青年麻醉 AI 助手（高级版）及 Gemini 2.0 pro 提出了允许性高碳酸血症，ChatGPT o1 提出了镇痛管理，Gemini 2.0 pro 提出了液体管理。这些内容都是对指南的有益补充。

2.2　LLM 评价及麻醉科医师同行评价的 TDS 与 QAMAI 评分　8 种 LLM 及麻醉科医师对 8 种模型评价的 TDS 见图 1。8 种模型 TDS 由低到高排序如下：ChatGPT o1［1.5（1.0，2.0）分］、DeepSeek R1［2.0（1.5，2.8）分］、豆包［2.0（1.2，3.8）分］、新青年麻醉 AI（高级版）［2.5（2.0，3.0）分］、文心大模型 4.0［2.5（2.0，3.8）分］、通义千问［2.5（2.0，4.5）分］、Gemini 2.0 pro［3.0（2.0，5.0）分］、Kimi［3.5（1.2，4.0）分］。各模型中位 TDS 差异无统计学意义（H＝10.428，P＝0.165 6），但 TDS 分布存在差异，ChatGPT o1 的 TDS 主要集中在 0～3 分的区间内且数据点相对集中，表明其评分较为稳定；Gemini 2.0 pro 的 TDS 主要集中在 2～5 分之间，而 Kimi 的评分则主要集中在 1～4 分之间，显示这 2 种模型在评分上存在一定的差异。

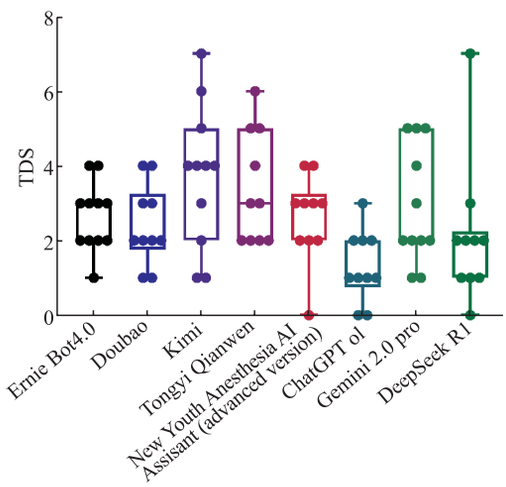

图 1　8 种 LLM 的 TDS 分布

Fig 1　Distribution for TDS of 8 types of LLMs

Kruskal-Wallis H test indicated no significant difference in TDS among the 8 LLMs (H＝10.428, P＝0.165 6). LLM: Large language model; TDS: Total disagreement score; AI: Artificial intelligence.

8 种 LLM 及麻醉科医师对 8 种模型评价的 QAMAI 评分见图 2。8 种模型 QAMAI 评分由高至低排序如下：ChatGPT o1［25.5（22.2，27.8）分］、DeepSeek R1［25.5（19.2，28.2）分］、豆包［22.5（20.2，26.8）分］、新青年麻醉 AI 助手（高级版）［22.5（20.0，25.5）分］、通义千问［22.5（15.8，24.8）分］、Kimi［22.0（20.0，25.5）分］、文心大模型 4.0［22.0（20.5，24.0）分］、Gemini 2.0 pro［21.0（16.8，23.8）分］。各模型的 QAMAI 评分差异无统计学意义（H＝3.972，P＝0.783 0），且评分均集中于 20～26 分，其中文心大模型 4.0 和 ChatGPT o1 的评分分布更集中，而豆包和通义千问的评分分布相对分散。ChatGPT o1 的 QAMAI 评分主要集中在 22～28 分，显示出其具有较好的性能。DeepSeek R1 的 QAMAI 评分在国内 LLM 中最高。Gemini 2.0 pro 的 QAMAI 评分波动较大，说明该模型争议较大。

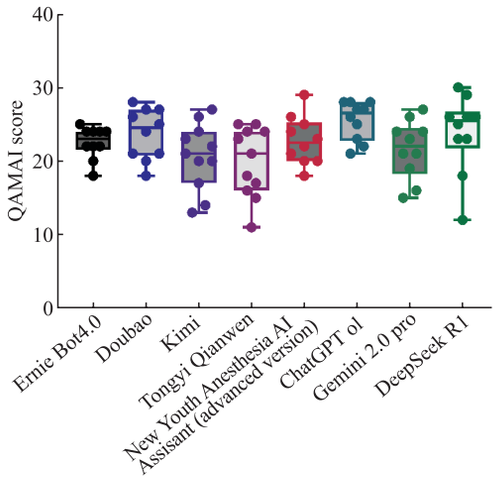

图 2　8 种 LLM 的 QAMAI 评分分布

Fig 2　Distribution for QAMAI scores of 8 types of LLMs

Kruskal-Wallis *H* test indicated no significant difference in QAMAI score among the 8 LMMs (H=3.972, P=0.783 0). LLM: Large language model; QAQMAI: Quality analysis of medical artificial intelligence; AI: Artificial intelligence.

2.3　各 LLM 评价与麻醉科医师同行评价的评分不符合比例及 Cliff's δ 效应量分析　8 种模型经各 LLM 评价 TDS 为不符合的比例由低到高排序为：ChatGPT o1（0/8）、DeepSeek R1（1/8）、新青年麻醉 AI 助手（高级版）（1/8）、文心大模型 4.0（2/8）、豆包（2/8）、通义千问（2/8）、Kimi（4/8）、Gemini 2.0 pro（4/8）；经各 LLM 评价 QAMAI 评分为不符合的比例由低到高排序为：ChatGPT o1（0/8）、新青年麻醉 AI 助手（高级版）（0/8）、文心大模型 4.0（0/8）、豆包（0/8）、DeepDeek R1（1/8）、Kimi（1/8）、通义千问（2/8）、Gemini 2.0 pro（2/8）。

Cliff's δ 的取值范围为−1～1，绝对值越大，表示两者之间的差异越大。由于 TDS 越低表示模型越接近指南，因此 Cliff's δ 负值表示 2 个模型中行模型优于列模型。各 LLM 的 Cliff's δ 效应量均值从小到大排序为：ChatGPT o1（−0.54）、DeepSeek R1（−0.18）、豆包（−0.10）、新青年麻醉 AI 助手（高级版）（−0.01）、文心大模型 4.0（0.14）、Kimi（0.16）、通义千问（0.21）、Gemini 2.0 pro（0.33）。ChatGPT o1 在 TDS 上的平均效应量为负值且绝对值最大，表示其偏离指南的程度最小，表现最优；而 Gemini 2.0 pro 的平均效应量为最大正值，表示其偏离指南的程度最大，表现相对较差。见表 2。

表 2　8 种大语言模型 TDS 的 Cliff's δ 效应量分析

Tab 2　Cliff's δ effect analysis of TDS for 8 large language models

δ

Large language model	Ernie Bot4.0	Doubao	Kimi	Tongyi Qianwen	New Youth Anesthesia AI Assistant (advanced version)	ChatGPT o1	Gemini 2.0 pro	DeepSeek R1
Ernie Bot4.0		0.219	−0.094	−0.062	0.125	0.656	−0.219	0.344
Doubao	−0.219		−0.219	−0.281	−0.062	0.406	−0.406	0.078
Kimi	0.094	0.219		−0.031	0.250	0.562	−0.250	0.266
Tongyi Qianwen	0.062	0.281	0.031		0.156	0.656	−0.047	0.344
New Youth Anesthesia AI Assistant (advanced version)	−0.125	0.062	−0.250	−0.156		0.484	−0.297	0.203
ChatGPT o1	−0.656	−0.406	−0.562	−0.656	−0.484		−0.688	−0.328
Gemini 2.0 pro	0.219	0.406	0.250	0.047	0.297	0.688		0.375
DeepSeek R1	−0.344	−0.078	−0.266	−0.344	−0.203	0.328	−0.375	

TDS: Total disagreement score; AI: Artificial intelligence.

QAMAI 评分越高表示模型越接近指南，因此 Cliff's δ 正值表示 2 个模型中行模型优于列模型。各 LLM 的 Cliff's δ 效应量均值从小到大排序为：ChatGPT o1（0.44）、DeepSeek R1（0.27）、豆包（0.08）、新青年麻醉 AI 助手（高级版）（−0.02）、Kimi（−0.09）、文心大模型 4.0（−0.14）、通义千问（−0.23）、Gemini 2.0 pro（−0.31）。ChatGPT o1 在 QAMAI 评分上的平均效应量最高，表示其符合指南的程度最大，表现最优；而 Gemini 2.0 pro 的平均效应量最低，表示其符合指南的程度最小，表现相对较差。见表 3。

LMM 评价结果表明 ChatGPT o1 生成的建议符合指南建议，多数 LLM 认为 Kimi、通义千问及 Gemini 2.0 pro 的建议不符合指南建议，其主要原因可能是其生成的部分建议严重偏离指南。例如，Kimi 未提供氧合指标且其肺复张策略建议的压力低于指南建议；通义千问提供的建议大多参考急性呼吸窘迫综合征的肺保护通气指南。

表 3　8 种大语言模型 QAMAI 评分的 Cliff's δ 效应量分析

Tab 3　Cliff's δ effect analysis of QAMAI scores for 8 large language models

δ

Large language model	Ernie Bot4.0	Doubao	Kimi	Tongyi Qianwen	New Youth Anesthesia AI Assistant (advanced version)	ChatGPT o1	Gemini 2.0 pro	DeepSeek R1
Ernie Bot4.0		−0.141	−0.016	0.016	−0.078	−0.578	0.234	−0.422
Doubao	0.141		0.172	0.297	0.078	−0.344	0.344	−0.141
Kimi	0.016	−0.172		0.109	−0.047	−0.500	0.172	−0.266
Tongyi Qianwen	−0.016	−0.297	−0.109		−0.172	−0.578	0.047	−0.453
New Youth Anesthesia AI Assistant (advanced version)	0.078	−0.078	0.047	0.172		−0.391	0.266	−0.234
ChatGPT o1	0.578	0.344	0.500	0.578	0.391		0.656	0.031
Gemini 2.0 pro	−0.234	−0.344	−0.172	−0.047	−0.266	−0.656		−0.422
DeepSeek R1	0.422	0.141	0.266	0.453	0.234	−0.031	0.422	

QAMAI: Quality assessment of medical artificial intelligence; AI: Artificial intelligence.

麻醉科医师同行评价 TDS 为不符合的比例由低到高的 LLM 排序为：ChatGPT o1（0/2）、DeepSeek R1（0/2）、文心大模型 4.0（0/3）、豆包（0/2）、Gemini 2.0 pro（0/3）、新青年麻醉 AI 助手（高级版）（1/2）、通义千问（2/2）、Kimi（3/3）。麻醉科医师同行评价 QAMAI 评分为不符合的比例由低到高的 LLM 排序为：ChatGPT o1（0/2）、DeepSeek R1（0/2）、文心大模型 4.0（0/2）、豆包（0/2）、Gemini 2.0 pro（0/2）、新青年麻醉 AI 助手（高级版）（0/2）、通义千问（2/3）、Kimi（2/3）。

综上可见，ChatGPT o1 和 DeepSeek R1 在 TDS 和 QAMAI 评分及不符合的比例方面表现最佳，而 Gemini 2.0 pro、Kimi 及通义千问表现相对较差。这些差异可能源于各模型在算法架构、训练数据时效性及医学逻辑推理能力等方面的异质性。

3　讨　论

在当今时代，LLM 正以前所未有的速度融入人们的工作和日常生活。在医疗领域，LLM 的辅助潜力备受关注。围手术期肺保护通气策略作为麻醉管理的关键环节，其具体方法在国内外相关指南中已有较为明确的阐述[10,16]。本研究旨在分析当前主流 LLM 提供的围手术期肺保护通气策略建议与指导临床实践的指南建议之间的契合度，从而探究 LLM 在该专业领域的应用价值。

本研究采用了 LLM 互评与麻醉科医师同行评议相结合的方式，尽管最终结果未呈现出统计学意义，但通过对各模型评分中位数的分析仍能发现不同模型之间的表现存在差异。其中，ChatGPT o1 在各项评估指标中表现突出，而 Kimi 和通义千问则相对一般。

本研究结果显示，虽然不同 LLM 所提供的肺保护通气策略在准确性上不存在显著差异，但从评分分布上看各模型之间存在明显不同。ChatGPT o1 能为围手术期肺保护通气策略提供较为全面的参考依据，国内 DeepSeek R1 也有出色表现。该结果表明，不同 LLM 在对互联网海量信息的检索和整合能力方面存在一定程度的差距。DeepSeek R1 的良好表现反映了国产 LLM 在医疗领域的广阔应用前景。专业模型新青年麻醉 AI 助手（高级版）的整体评分处于中等水平，但在通气目标具体数值上具有较高准确性，体现出该模型在专业领域的独特优势和一定实用价值。

ChatGPT o1 提供的内容基本覆盖指南所提及的各项建议，且在具体数据方面与指南高度契合。而 Gemini 2.0 pro、Kimi 和通义千问存在明显不足，在关键参数上这几个模型给出的建议与指南差异较大，在策略的全面性上也有所欠缺。Gemini 2.0 pro 和 Kimi 仅对麻醉期间的肺保护通气策略提出建议，忽视了围手术期其他阶段的管理；通义千问虽涉及术后管理，但描述过于简略，且对存在特定疾病的患者考虑不周全，难以满足复杂多变的临床实际需求。这反映出部分 LLM 在处理专业医学问题时不够精细和全面，在应用于临床实践时需要谨慎评估。

各模型生成的建议在整体方向上与指南保持基本一致，这无疑展示了 LLM 在医学知识整合方面的积极成果。但是没有任何一个模型的建议能够与指南达到完全契合的程度。因此，在看待 LLM 在医疗领域的应用时必须保持理性和审慎态度。临床医师应将 LLM 的建议与自身临床经验、指南相结合以进行全面综合考量和判断，确保医疗决策的科学性、安全性和有效性。此外，未来研究可进一

步深入探索 LLM 在临床实践中的具体应用场景。例如，在术前评估阶段，ChatGPT o1 和 DeepSeek R1 可辅助麻醉医师根据患者的具体情况制定个体化通气策略；在术中管理阶段，它们可帮助医生实时监测患者的生理指标，如气道压力、潮气量等，并提供即时的通气参数调整建议，以确保患者的安全；在术后恢复阶段，这 2 个模型则可帮助医生评估患者的恢复情况，预测可能出现的并发症，并制定相应的康复计划。为了全面验证 LLM 生成的建议在实际临床环境中的有效性和安全性，未来研究应积极开展多中心、大样本的临床研究，为其在临床实践中的广泛应用提供坚实的证据支持。

本研究也有不足之处。围手术期肺保护通气策略的指南和专家共识众多，不同指南的具体意见也存在部分差异，本研究仅选取其中一项作为评分的参考依据，这可能对评分的准确性产生影响。此外，在处理 LLM 给出的建议时仅采用其提供的首个答案，并未对模型进行反复训练。虽然该方式能直观展现模型的检索整合能力，却忽略了其学习能力，由于采用不同的算法架构，不同 LMM 所给出的答案也会因此产生偏差。最后，尽管 TDS 及 QAMAI 评分在人工智能相关研究中应用广泛，其评分标准仍存在一定主观性。不同评分者对于评分标准的理解和把握也可能存在差异，这种理解上的偏差可能导致评分结果的不一致。因此，对于当前 LLM 建议及评分的结论仍需保持谨慎态度。鉴于人工智能技术相关研究的迅猛发展之势，这一结论或许会随着技术进步及评分标准的更新而迅速改变。

［参 考 文 献］

［1］ 中国医师协会麻醉学医师分会，中国心胸血管麻醉学会胸科麻醉分会，中国心胸血管麻醉学会麻醉与身心医学分会. 急诊手术患者围术期肺保护管理策略的专家共识（2024 版）[J]. 中华麻醉学杂志，2025，45（1）：31-41. DOI: 10.3760/cma.j.cn131073-20241201-00108.

［2］ ODOR P M, BAMPOE S, GILHOOLY D, et al. Perioperative interventions for prevention of postoperative pulmonary complications: systematic review and meta-analysis[J]. BMJ, 2020, 368: m540. DOI: 10.1136/bmj.m540.

［3］ SHAH N H, ENTWISTLE D, PFEFFER M A. Creation and adoption of large language models in medicine[J]. JAMA, 2023, 330(9): 866-869. DOI: 10.1001/jama.2023.14217.

［4］ 熊利泽. 2025：争做有“人工智能”素养的麻醉人[J]. 中华麻醉学杂志，2025，45（1）：1-2. DOI: 10.3760/cma.j.cn131073-20250107-00101.

［5］ LOPES S, ROCHA G, GUIMARÃES-PEREIRA L. Artificial intelligence and its clinical application in anesthesiology: a systematic review[J]. J Clin Monit Comput, 2024, 38(2): 247-259. DOI: 10.1007/s10877-023-01088-0.

［6］ LONSDALE H, BURNS M L, EPSTEIN R H, et al. Strengthening discovery and application of artificial intelligence in anesthesiology: a report from the Anesthesia Research Council[J]. Anesthesiology, 2025, 142(4): 599-610. DOI: 10.1097/ALN.0000000000005326.

［7］ ABDEL MALEK M, VAN VELZEN M, DAHAN A, et al. Generation of preoperative anaesthetic plans by ChatGPT-4.0: a mixed-method study[J]. Br J Anaesth, 2025, 134(5): 1333-1340. DOI: 10.1016/j.bja.2024.08.038.

［8］ LU Y, ALETA A, DU C, et al. LLMs and generative agent-based models for complex systems research[J]. Phys Life Rev, 2024, 51: 283-293. DOI: 10.1016/j.plrev.2024.10.013.

［9］ SAXENA S, BARRETO CHANG O L, SUPPAN M, et al. A comparison of large language model-generated and published perioperative neurocognitive disorder recommendations: a cross-sectional web-based analysis[J]. Br J Anaesth, 2025: S0007-S0912(25)00006-6. DOI: 10.1016/j.bja.2025.01.001.

［10］ YOUNG C C, HARRIS E M, VACCHIANO C, et al. Lung-protective ventilation for the surgical patient: international expert panel-based consensus recommendations[J]. Br J Anaesth, 2019, 123(6): 898-913. DOI: 10.1016/j.bja.2019.08.017.

［11］ 薄禄龙，卞金俊，邓小明，等. 手术患者肺保护性通气策略：国际专家组推荐规范的解读[J]. 国际麻醉学与复苏杂志，2020，41（5）：417-421. DOI: 10.3760/cma.j.cn321761-20200110-00020.

［12］ SAIBENE A M, ALLEVI F, CALVO-HENRIQUEZ C, et al. Reliability of large language models in managing odontogenic sinusitis clinical scenarios: a preliminary multidisciplinary evaluation[J]. Eur Arch Otorhinolaryngol, 2024, 281(4): 1835-1841. DOI: 10.1007/s00405-023-08372-4.

［13］ VAIRA L A, LECHIEN J R, ABBATE V, et al. Validation of the quality analysis of medical artificial intelligence (QAMAI) tool: a new tool to assess the quality of health information provided by AI platforms[J]. Eur Arch Otorhinolaryngol, 2024, 281(11): 6123-6131. DOI: 10.1007/s00405-024-08710-0.

［14］ ALLISON S L, KOSCIK R L, CARY R P, et al. Comparison of different MRI-based morphometric estimates for defining neurodegeneration across the Alzheimer’s disease continuum[J]. Neuroimage Clin, 2019, 23: 101895. DOI: 10.1016/j.nicl.2019.101895.

［15］ LOOS N L, SELLES R W, TER STEGE M H P, et al. Using outcome information during consultation yields better shared decision making, better patient experiences, and more positive expectations: a comparative effectiveness study[J]. Value Health, 2025, 28(4): 571-581. DOI: 10.1016/j.jval.2025.01.009.

［16］ 中华医学会麻醉学分会“围术期肺保护性通气策略临床应用专家共识”工作小组. 围术期肺保护性通气策略临床应用专家共识[J]. 中华麻醉学杂志，2020，40（5）：513-519. DOI: 10.3760/cma.j.cn131073.20200402.00501.

［本文编辑］　杨亚红