

人工智能在复杂临床诊断中的应用： 一例以隐匿性记忆力减退为首发表现的 中枢神经系统生殖细胞瘤

——基于大语言模型在多阶段复杂临床诊疗中的推理能力系统评估

高同胜^{1,2}, 谢晓成³, 舒徐凯⁴, 付应勤³, 黄辉敏⁴, 王 晓^{5*}, 吕海龙^{5*}

¹宁波大学医学部, 浙江 宁波

²金华市第二医院心理科, 浙江 金华

³瑞安市第五人民医院三病区, 浙江 温州

⁴温州医科大学附属第三医院精神卫生科, 浙江 温州

⁵浙江大学医学院附属第一医院精神卫生科, 浙江 杭州

收稿日期: 2026年7月18日; 录用日期: 2026年8月11日; 发布日期: 2026年8月19日

摘 要

目的: 中枢神经系统生殖细胞瘤(CNS-GCTs)极少表现为单纯性认知症状, 临床上极易导致漏诊与误诊。本文报道了一例相关罕见病例, 并借此评估了五种人工智能(AI)大语言模型的临床诊断推理能力。方法: 患者为16岁男性, 因“进行性记忆力减退一年”就诊, 初步检查提示严重高钠血症, 最终确诊为鞍上中枢神经系统生殖细胞瘤。本研究将该病例三次就诊阶段的临床资料分别输入五种AI模型中, 并由20名精神科医生采用修订版IDEA评分量表(Revised IDEA Score)对模型生成的推理过程进行量化评分。结果: 放疗后患者肿瘤完全消退, 但记忆损害表现持续存在。在AI模型评估中, 尽管有三款模型展现出较高且稳定的临床推理质量, 但整体诊断准确率仅为40.0%, 最终仅有两款模型正确诊断出该肿瘤。结论: 对于出现隐匿性认知功能减退的青少年患者, 临床应高度警惕CNS-GCTs的可能。虽然先进的AI能够模拟高水平的临床推理, 但在面对非典型病例时, 其诊断准确率仍存在局限性。这表明在“复制推理逻辑”与“确保诊断正确”之间, AI仍存在关键技术鸿沟。未来的研究需进一步深挖人工智能在医疗场景中的应用, 以期实现高质量、高精度的临床决策。

关键词

中枢神经系统生殖细胞瘤, 记忆力减退, 人工智能, 临床推理

*共同通讯作者。

Applying Artificial Intelligence to a Complex Diagnostic Challenge: A CNS Germ Cell Tumor Presenting as Insidious Memory Loss

—A Systematic Evaluation of Large Language Models' Reasoning Abilities across Multi-Stage Complex Clinical Workups

Tongsheng Gao^{1,2}, Xiaocheng Xie³, Xukai Shu⁴, Yingqin Fu³, Huimin Huang⁴, Xiao Wang^{5*}, Hailong Lyu^{5*}

¹Health Science Center, Ningbo University, Ningbo Zhejiang

²Department of Psychiatry, The Second Hospital of Jinhua, Jinhua Zhejiang

³Psychiatric Unit 3, Ruian Fifth People's Hospital, Wenzhou Zhejiang

⁴Department of Psychiatry, The Third Affiliated Hospital of Wenzhou Medical University, Wenzhou Zhejiang

⁵Department of Psychiatry, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou Zhejiang

Received: July 18, 2026; accepted: August 11, 2026; published: August 19, 2026

Abstract

Objective: Central nervous system germ cell tumors (CNS-GCTs) rarely present with isolated cognitive symptoms, causing significant diagnostic delays. We report such a case to evaluate the diagnostic reasoning of five leading artificial intelligence (AI) large language models. **Methods:** A 16-year-old male with a one-year history of progressive memory impairment was diagnosed with a suprasellar CNS-GCT after a workup revealed severe hyponatremia. Clinical data from three diagnostic stages were provided to five AI models. Twenty psychiatrists scored the AI-generated reasoning using the validated the Revised IDEA Score. **Results:** After radiotherapy, the patient's tumor resolved, but his memory impairment persisted. In the AI evaluation, while three of the five models consistently demonstrated high-quality clinical reasoning, the overall diagnostic accuracy was only 40.0%, with just two models correctly identifying the tumor. **Conclusion:** CNS-GCTs should be suspected in adolescents with insidious cognitive decline. Although advanced AI can emulate high-level clinical reasoning, its diagnostic accuracy in atypical cases remains unreliable. This demonstrates a critical gap between replicating a reasoning process and achieving diagnostic certainty. Future research should further explore the specific applications of artificial intelligence in medical settings to enable high-quality and highly accurate clinical decisions.

Keywords

Central Nervous System Germ Cell Tumors, Memory Loss, Artificial Intelligence, Clinical Reasoning

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

中枢神经系统生殖细胞瘤(Central Nervous System Germ Cell Tumors, CNS-GCTs)是一类罕见的恶性肿瘤,多发于儿童及青少年群体[1]。其临床表现因肿瘤浸润部位而异,最常累及的部位为松果体区(约占 50%)和鞍上区(约占 20%~30%)。其中,松果体区肿瘤常因压迫导致脑积水,引发颅内压增高,并可伴发帕里诺综合征(Parinaud syndrome) [2]。相比之下,精神症状属于 CNS-GCTs 的非典型表现。目前仅有少数文献报道了 CNS-GCTs 伴发精神症状的病例,临床表型包括精神运动性迟滞、行为与情绪改变、言语障碍以及认知功能损害等[3]。然而,以精神症状为首发表现的 CNS-GCTs 因病程隐匿且涉猎多学科交叉,极易导致临床漏诊与延误诊断。

人工智能(Artificial Intelligence, AI)是指利用计算机对人类智能进行模拟的技术。自 20 世纪 50 年代诞生以来, AI 技术历经持续演进,如今已步入以生成式人工智能为代表的 3.0 时代[4]。目前,人工智能、机器学习及深度学习已广泛应用于医疗健康领域,涵盖临床决策支持系统、实验室诊断、临床影像学分析以及公共卫生疫情管理等多个方面[5]。既往研究表明, AI 能够从多维度赋能肿瘤的全程管理,例如精准评估肿瘤总体体积、预测临床结局以及评估疾病预后影响等[6]。在人工神经网络技术的支撑下,深度学习已在常规核医学、单光子发射计算机断层显像(Single-Photon Emission Computed Tomography, SPECT)、正电子发射断层显像(Positron Emission Tomography, PET)、计算机断层扫描(Computed Tomography, CT)及磁共振成像(Magnetic Resonance Imaging, MRI)中得到了深度融合与应用[7]。同时,多项研究证实, AI 在精神心理健康领域的病情筛查、疾病分类、疗效预测及随访监测中均表现出良好的应用价值[8]。亦有学者发现, AI 结合机器学习在优化临床决策流程方面蕴含巨大潜力,能够通过系统性分析患者的完整病史、实验室检验及影像学检查结果,提供可靠的鉴别诊断建议[9]。

本研究旨在报道一例以隐匿性记忆力减退为首发症状的青少年 CNS-GCTs 罕见病例。我们将这一复杂的临床真实案例输入五款人工智能大语言模型中进行推理分析,系统性评估了 AI 在面对非典型临床表现时对罕见病的诊断推理能力。本研究期望能为人工智能在神经肿瘤学与精神医学交叉领域的应用中提供新的视角与实证依据。

2. 病例汇报

患者为 16 岁男性,既往生长发育史无异常。入院前一年,患者逐渐出现记忆力减退,严重影响其学习能力。升入高中后,上述症状进一步加重,表现为无法进行复杂的数学运算。入院前六个月,患者突发顺行性遗忘,无法回忆前一天自己的言行。一周后,父母带其至我院就诊。

体格检查显示患者面色苍白,皮肤干燥。精神检查过程中,患者表现为不合作、反应迟钝,且无法回忆前一天发生的事情。心理测量评估结果显示:焦虑自评量表(Self-Rating Anxiety Scale, SAS)评分为 43 分,抑郁自评量表(Self-Rating Depression Scale, SDS)评分为 54 分,心境障碍问卷(Mood Disorder Questionnaire, MDQ)评分为 1 分,轻躁狂检测清单(Hypomania Checklist-32, HCL-32)评分为 5 分,提示存在轻度焦虑与轻度抑郁状态。同时,患者完善了相关实验室检查,包括甲状腺功能、肿瘤标志物、电解质、性激素及血清皮质醇水平等。结果显示:血清钠水平极高(171 mmol/L),血清氯水平轻度升高(129 mmol/L),催乳素(92.45 ng/ml)和总甲状腺素(50.26 nmol/l)水平显著升高。颅脑 MRI 提示鞍上池前部见混合信号影,双侧侧脑室室管膜增厚。

入院第四天,经神经外科专家会诊后,患者转诊以进一步排查中枢神经系统生殖细胞瘤。此时,患者的血清钠、血清氯、催乳素及总甲状腺素水平仍显著增高;胰岛素样生长因子-1(41.9 ng/l)、人绒毛膜促性腺激素(33.9 mIU/ml)以及血清渗透压(327 mOsm/kg)均显著升高。随后行腰椎穿刺术,脑脊液

(Cerebrospinal Fluid, CSF)检查显示：颅内压为 190 mm H₂O，潘氏试验(Pandy test)阴性(-)，肿瘤标志物阴性(-)，未见脱落细胞(-)；氯 144 mmol/L (↑)，葡萄糖 3.2 mmol/L，蛋白 0.86 mg/dL (↑)，有核细胞计数 20/uL，红细胞计数 25/uL。颅脑增强 MRI 提示：鞍上池见一囊实性结节，累及垂体柄、漏斗部及视神经孔。至此，患者最终确诊为 CNS-GCTs。鉴于肿瘤生长部位手术难度极大，患者被转至上级医院接受了 18 次近距离内照射放疗和 7 次外照射放疗。放疗结束四个月后复查颅脑 MRI，结果提示肿瘤病灶完全消退，血清钠水平稳定在 146~151 mmol/L；然而，在最后一次随访中，家属主诉患者仍存在记忆障碍，且伴有意志减退、精神萎靡，至今未能复学。

本例患者的诊断过程展现了三阶段临床诊断推理特征，详见图 1。

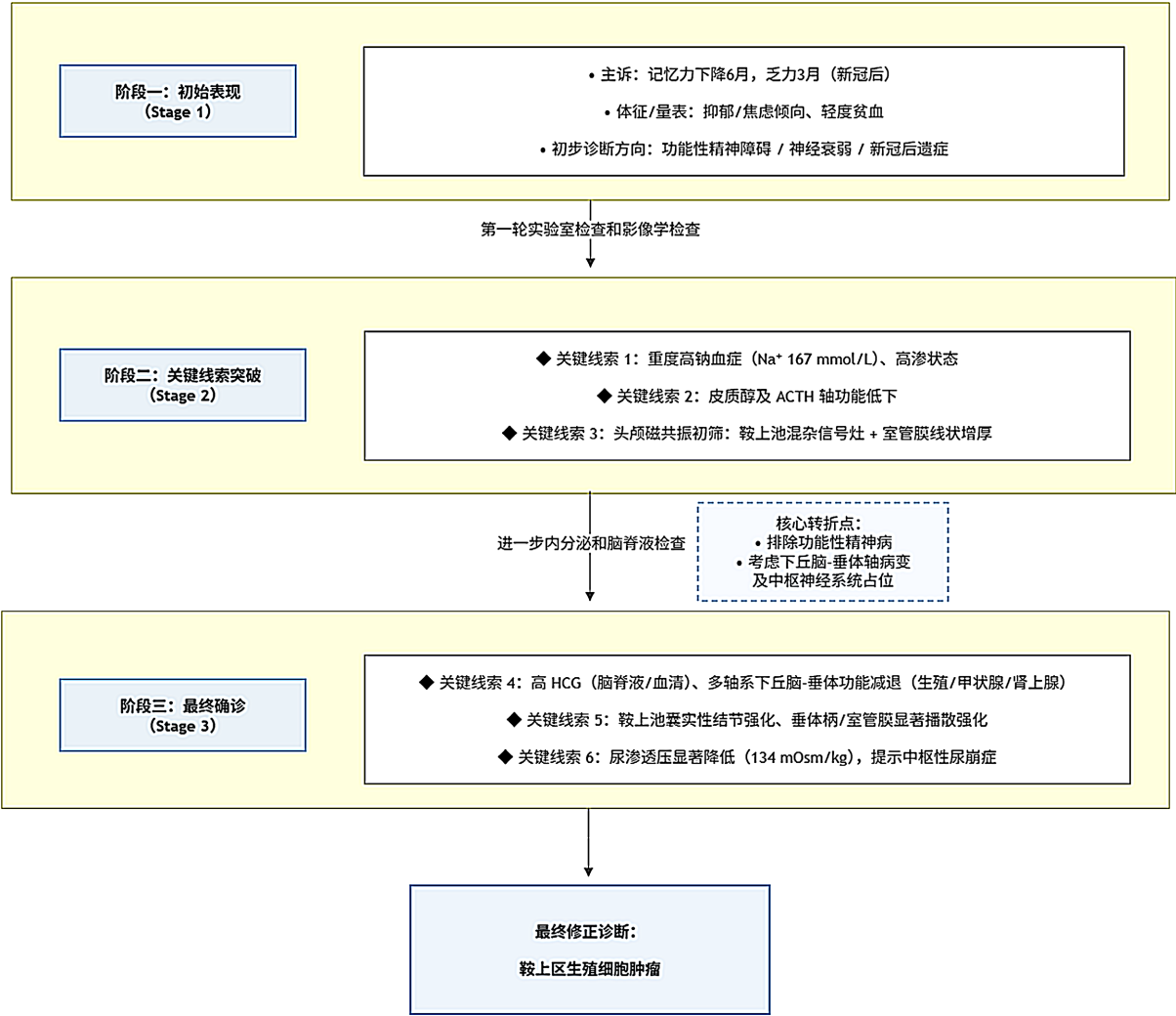


Figure 1. Three-stage clinical evolution and diagnostic reasoning flowchart based on the patient's clinical data
图 1. 基于患者临床数据的三阶段临床演变与诊断推理流程图

3. 材料与方法

3.1. 研究设计与流程

为了模拟真实的临床诊疗过程，本研究将按照时间顺序将该患者所有临床资料划分为三个不同的就

诊阶段。该设计旨在精确还原患者在医疗系统中的就诊轨迹：(1) 门诊就诊阶段：第一组数据集包含了患者首次就诊于精神科门诊时的所有临床资料；(2) 住院收治阶段：第二组数据集为累加数据，加入了患者收入精神科病房后完善的各项新检查结果；(3) MDT 会诊阶段：第三组亦是最终的数据集，包含了在最终确诊的多学科联合会诊(Multidisciplinary Team, MDT)前所能获取的全部临床信息。然后，本研究将国内外主流的五款人工智能模型——Grok3、DeepSeek-R1、Gemini 2.0、ChatGPT-O 和 Claude 3.7 纳入评估过程。

在每个阶段中，研究人员将包含相应临床数据(病历记录、体格检查/神经系统检查、实验室检验以及神经心理学评估)的标准提示词，依序输入到每款 AI 模型中。每款模型在每个阶段各生成一份临床推理答复，共计获得 15 份独立答复以供评估。随后，由 20 名精神科医生采用修订版 IDEA 评分量表(Revised IDEA Score)对这些答复进行独立量化评分[10]。该量表涵盖四个核心维度的评估：叙述性总结(Interpretive summary)、鉴别诊断(Differential diagnosis)、主诉诊断依据(Explanation of lead diagnosis)以及备选诊断(Alternative diagnoses)。总分超过 6 分即判定为高质量临床推理。具体的实验流程见图 2。

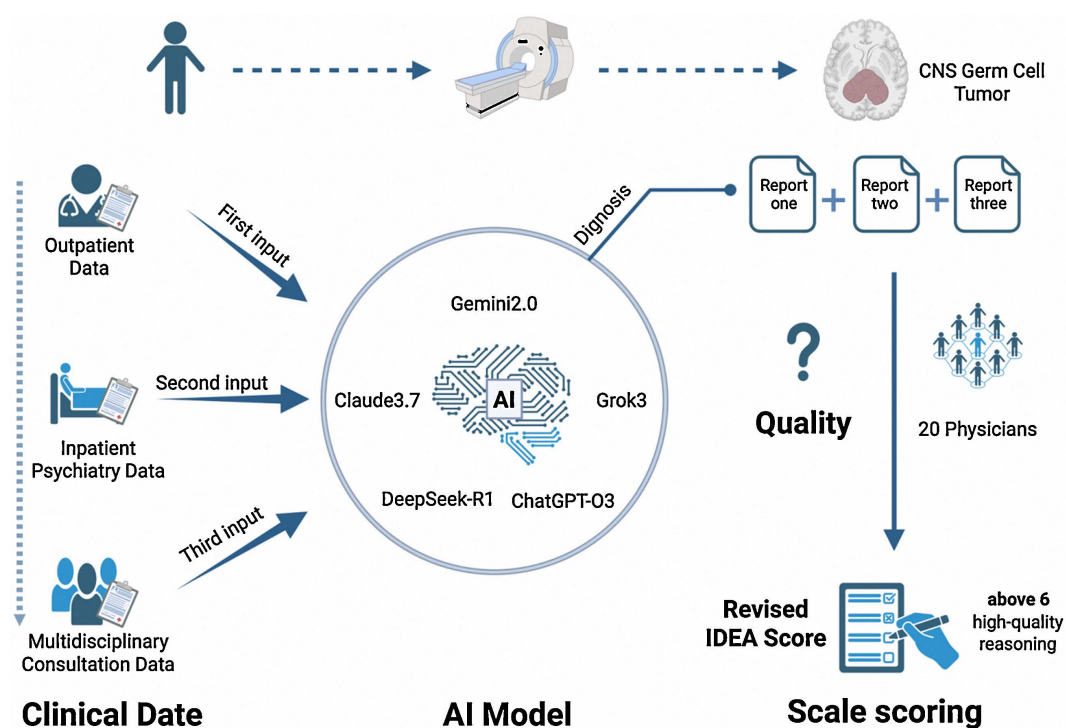


Figure 2. Research steps to use five AI models for clinical reasoning and diagnosis based on clinical data

图 2. 基于临床数据利用五款 AI 模型进行临床推理与诊断的研究流程图

3.2. 统计学分析

常规统计学分析均采用 R 4.3.3 软件(<https://www.r-project.org/>)进行。统计描述指标包括均值(Mean)、四分位数(Quartiles: Q1, Q2, Q3)以及众数(Mode)。

4. 结果

4.1. 五款 AI 模型修订版 IDEA 评分的基础统计学分析

20 名精神科医生分别对 15 份答复进行了修订版 IDEA 评分量表测评。基础统计学结果见表 S1 和图 3。输入临床资料后，各模型在四个维度的平均分按降序排列，具体表现如下：

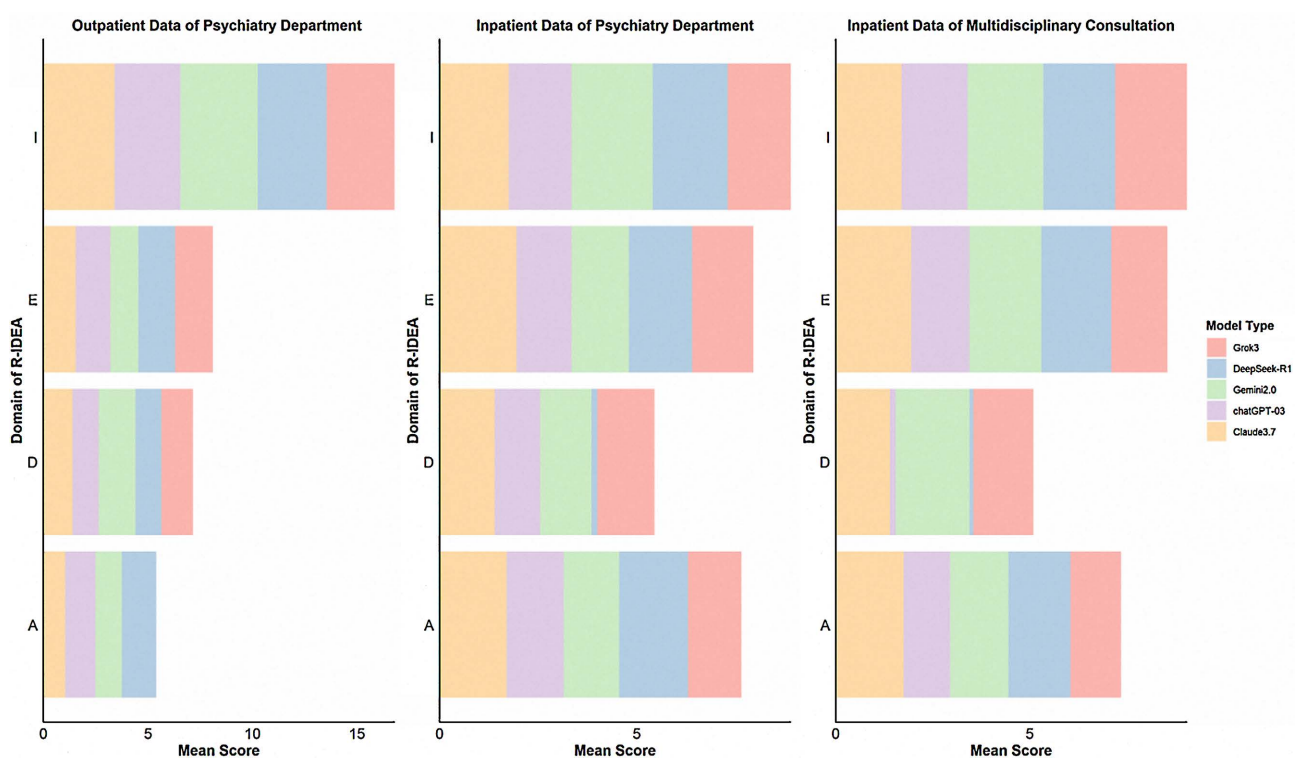


Figure 3. Mean scores of different models across domains of R-IDEA based on medical records of three visits respectively
图 3. 分别基于三次就诊病历的 R-IDEA 各维度不同模型平均得分

- (1) 门诊阶段数据:
- 领域 I(叙述性总结): Gemini 2.0 得分最高; ChatGPT-O3 得分最低。
- 领域 D(鉴别诊断): Gemini 2.0 表现最具优势; ChatGPT-O3 与 DeepSeek-R1 得分垫底。
- 领域 E(主诉诊断依据): Grok3 居于首位; Gemini 2.0 则处于劣势。
- 领域 A(备选诊断): DeepSeek-R1 得分最高; Grok3 得分最低。
- (2) 精神科住院阶段数据:
- 领域 I(叙述性总结): Gemini 2.0 得分最高; ChatGPT-O3 得分最低。
- 领域 D(鉴别诊断): Grok3 居于首位; DeepSeek-R1 得分最低。
- 领域 E(主诉诊断依据): Claude 3.7 表现最具优势; ChatGPT-O3 得分垫底。
- 领域 A(备选诊断): DeepSeek-R1 得分最高; Grok3 得分最低。
- (3) 多学科联合会诊(MDT)阶段数据:
- 领域 I(叙述性总结): Gemini 2.0 得分最高; ChatGPT-O3 得分最低。
- 领域 D(鉴别诊断): Gemini 2.0 居于首位; DeepSeek-R1 得分最低。
- 领域 E(主诉诊断依据): Claude 3.7 表现最具优势; Grok3 得分垫底。
- 领域 A(备选诊断): Claude 3.7 得分最高; Grok3 得分最低。

4.2. 五款 AI 模型修订版 IDEA 总分对比

五款 AI 模型的平均总分如表 S1 和图 4 所示。

门诊阶段数据: 所有模型的评分均超过了 6 分的临床高质量分水岭, 其中 Gemini 2.0 的得分最高(8.05 分)。

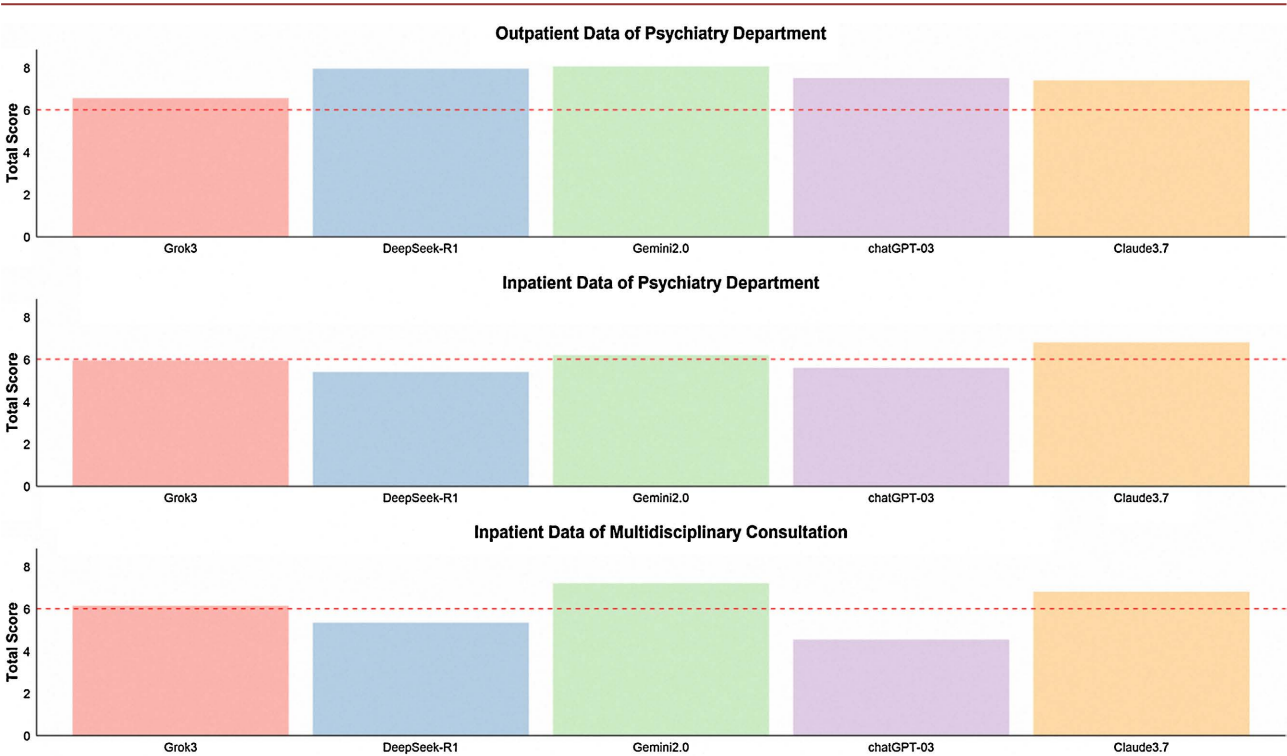


Figure 4. Total scores of different models of R-IDEA based on medical records of three visits respectively

图 4. 分别基于三次就诊病历的 R-IDEA 各模型总分对比

精神科住院阶段数据：仅有 Claude 3.7、Gemini 2.0 和 Grok3 的总分保持在 6 分以上，其中 Claude 3.7 的表现最佳(6.80 分)。

多学科联合会诊(MDT)阶段数据：同样仅有 Claude 3.7、Gemini 2.0 和 Grok3 的总分超过 6 分，此阶段 Gemini 2.0 的综合表现最好(7.20 分)。

4.3. 五款 AI 模型的诊断准确率

五款 AI 模型给出的最终临床诊断详见表 1。在所有模型中，仅有 DeepSeek-R1 和 Claude 3.7 准确识别出了“鞍区生殖细胞瘤”或“颅内生殖细胞瘤”，整体诊断准确率仅为 40.0%。

Table 1. Diagnostic accuracy rates of different models

表 1. 不同模型诊断准确率

模型	最终诊断	准确率
Grok 3	垂体功能减退症	0
DeepSeek-R1	鞍区生殖细胞肿瘤	100%
Gemini 2.0	颅咽管瘤	0
ChatGPT-o3	垂体腺瘤	0
Claude 3.7	颅内生殖细胞肿瘤	100%
总计	/	40.00%

注：Grok 3 为 Grok 第三代模型；DeepSeek-R1 为 DeepSeek 的 R1 模型；Gemini 2.0 为 Gemini 2.0 Flash Thinking 模型；ChatGPT-o3 为 ChatGPT 的 o3 模型；Claude 3.7 为 Claude 3.7 Sonnet 模型。

4.4. Claude 与 ChatGPT-O3 的文本推理比较

为进一步探讨不同大语言模型诊断表现差异的潜在原因，本研究进一步比较了 Claude 3.7 与 ChatGPT-O3 在连续诊断阶段中的推理过程(详见表 2)。结果显示，两种模型在初诊阶段均考虑精神心理因素，但随着高钠血症、皮质醇降低及鞍上区异常影像学证据的出现，Claude 3.7 能够进一步整合下丘脑 - 垂体轴异常线索，而 ChatGPT-O3 仍倾向于将异常表现归因于病毒感染后综合征。

Table 2. Comparison of diagnostic reasoning trajectories generated by Claude and ChatGPT-O3 large language models in a case of CNS-GCTs

表 2. Claude 与 ChatGPT-O3 大语言模型在鞍上生殖细胞瘤诊断过程中的诊断推理比较

临床阶段	关键临床信息	Claude3.7 大模型的外显推理文本	ChatGPT-O3 大模型的外显推理文本	逻辑链比较
阶段 1: 门诊就诊阶段	16 岁男性，近记忆下降 6 个月，乏力 3 个月，病毒感染后加重，SDS 54 分	“患者存在情绪低落、兴趣减退、注意力下降，但同时需考虑甲状腺功能减退、营养不良性贫血及病毒感染后遗症。”	“患者表现符合抑郁症，SDS 评分提示中度抑郁；同时考虑慢性疲劳综合征和病毒感染后遗症。”	两模型均识别精神症状。但 Claude 同时保留躯体疾病鉴别，而 ChatGPT-O3 更倾向将症状归入精神/感染后状态，存在早期诊断收敛倾向。
阶段 2: 住院收治阶段	Na ⁺ 167 mmol/L，高渗状态，皮质醇显著降低，MRI 鞍上异常	“鞍上池异常信号提示垂体或下丘脑病变；低皮质醇及电解质异常提示内分泌轴受累。”	“高钠可能提示脱水或肾上腺功能异常；MRI 异常可能为炎症改变或病毒感染后神经系统影响。”	Claude 完成“电解质异常→内分泌异常→结构病变”的整合；ChatGPT-O3 虽然识别异常，但仍受病毒感染背景影响，对器质性病变解释不足。
阶段 3: MDT 会诊	睾酮降低、LH/FSH 降低、PRL 升高、血高渗低尿渗、HCG 升高、鞍上占位	“多轴垂体功能减退、中枢性尿崩以及鞍上区病变符合颅内生殖细胞瘤表现。”	“高催乳素、低睾酮及垂体柄增粗提示垂体腺瘤或其他垂体病变；同时考虑病毒感染后遗症。”	Claude 识别 HCG 升高这一关键肿瘤标志物，并建立影像 - 内分泌 - 肿瘤标志物闭环；ChatGPT-O3 更依赖常见垂体疾病模式，未充分利用特异性指标。

5. 讨论

本研究报道了一例以记忆力减退为首表现的罕见中枢神经系统生殖细胞瘤病例，并首次利用五款人工智能大语言模型对其临床诊疗数据进行了系统性分析。结果显示，Claude 3.7、Gemini 2.0 和 Grok3 展现出了高质量的临床推理能力，而 DeepSeek-R1 和 Claude 3.7 则实现了准确的临床诊断。从各评价维度来看，Gemini 2.0 在叙述性总结和鉴别诊断方面表现优异，Claude 3.7 擅长阐述主诉诊断依据，而 DeepSeek-R1 则在备选诊断的推导中更具优势。

本例患者在临床上主要表现为记忆力减退、高钠血症及垂体功能减退。尽管过往文献认为高钠血症可能通过破坏血脑屏障或引发脱髓鞘病变等机制导致认知功能障碍[11]，但目前的临床研究结论尚存争议。部分研究指出，血清钠水平与认知功能呈负相关[12]；另有研究则显示，高钠血症与认知障碍之间并无直接关联，反而是低钠血症与记忆功能损害密切相关[13]。同时，患者的记忆损害可能与其肿瘤侵犯下丘脑 - 边缘系统(如乳头体、穹窿及 Papez 环路等核心记忆结构)相关[14]。既往研究表明，鞍上区占位累及上述通路可导致持续性遗忘；而下丘脑 - 垂体轴激素缺失亦会通过削弱海马神经可塑性，加重认知与记忆功能障碍[15] [16]。

目前，CNS-GCTs 的临床确诊手段已较为成熟，且通过规范化治疗通常能取得较满意的疗效[17]。一项针对 455 例 CNS-GCTs 患者的长期随访研究显示，患者的 20 年和 30 年生存率分别高达 84.1%和 61.9%

[18]。然而,另一项研究发现,尽管有7例CNS-GCTs患者在治疗后得以长期存活,但其中5例在术后仍存在持续的精神症状且未能恢复[3]。这些证据提示,对于以精神症状为首发表现的CNS-GCTs患者,早期确诊可能直接关系到其精神心理预后的改善。这一观点也在本案患者身上得到了印证:该患者虽然在接受标准放疗后肿瘤临床缓解,但其短期的记忆损害在随访中仍未见明显恢复。

本研究结果表明,Claude 3.7、Gemini 2.0 和 Grok3 这三款AI模型能够输出高质量的临床推理。这一发现与既往研究相契合。此前已有数项研究尝试利用自然语言处理和人工智能技术来优化临床诊疗流程[19]。有学者证实,AI能够通过系统性整合患者的体征数据与临床信息,生成个体化的治疗方案[20]。同时,本研究也证实了两款AI模型在面对这一复杂疾病时能够给出精准诊断。既往文献同样指出,AI在破解罕见病领域“确诊率低、患者基数小”等诊断技术瓶颈方面蕴含巨大潜力[21]。目前,AI已被拓展用于肿瘤患者的个体化风险评估与预后康复,比如Salome等基于AI实现了癌症相关认知损害风险的早期识别;Svyat等则利用AI成功预测了脑肿瘤患者的死亡率、神经功能状态及治疗结局[22][23]。尽管临床医生普遍对AI辅助决策持积极态度,但对于AI的推理逻辑与人类临床思维之间的差异仍存顾虑[24]。目前多项研究正持续关注AI技术的成熟度。在真实的临床医疗场景中,其落地应用仍受限于技术与人为的多重因素,包括数据质量瓶颈、数据集漂移以及自动化偏见等[25]。大语言模型的输出差异源于其训练语料分布与知识边界。受限于数据偏倚、知识缺口及概率推理机制,模型易生成流畅却偏离事实的“幻觉”。目前医疗大预言模型的评价体系多局限于诊断准确率,对其模型偏倚、公平性及不确定性校准能力的评估依然匮乏[26][27]。在精神医学领域,早期研究还警告称,生成式AI在数据安全和患者隐私保护方面仍面临严峻的技术挑战[28]。随着AI在医学领域的深度渗透,科研人员对其技术局限性与医学伦理边界仍需保持高度审慎。

当然,本研究也存在一定的局限性。首先,作为单病例报告,本研究结论的普适性受到根本性限制。该患者极其非典型的临床表现具有特殊性,因此AI模型在此单一特殊病例上的特定表现,可能无法代表其在普遍临床场景下的真实水平,未来仍需开展多中心、大样本的病例系列研究加以验证。其次,患者持续性记忆损害的病因因果关系仍不明朗。目前尚难以厘清究竟是早期肿瘤压迫导致的器质性损伤,还是后期放疗带来的潜在神经毒性反应所致。要明确这一问题,需要配合标准化的纵向神经心理学评估进行长期随访。最后,本研究对AI能力的评估方法学仍存在一定局限。评估结果对提示词的设计具有较高的敏感性,且本研究未能报告专家评分者之间的评分一致性。未来的研究应当建立更加标准化的AI医疗评估框架,并常规报告组内相关系数等一致性信度指标,以确保研究结论的可靠性。

6. 结论

中枢神经系统生殖细胞瘤在发病初期可能仅表现为记忆力减退等隐匿性认知症状,临床上极易导致严重的漏诊与延误诊断。尽管当前最先进的人工智能(Artificial Intelligence, AI)模型已经能够复刻出高质量的临床推理逻辑,但在面对此类非典型、复杂的临床罕见案例时,其诊断准确率仍显不足。这表明,虽然人工智能可以作为强有力的辅助工具,有效帮助临床医生提高诊疗决策效率,但最终确诊仍高度依赖于临床医生丰富的临床经验与综合判断。

致 谢

感谢患者对本研究的参与,并衷心感谢20位精神科同行对本研究工作给予的大力支持。

声 明

本研究已获得浙江大学医学院附属第一医院医院伦理委员会的批准,批件号:Reference Number: 浙

大一院伦审 2023 研第 0582 号-快。鉴于本研究属于病例报告，已依规提交免除患者知情同意的申请。

基金项目

本研究获得浙江省自然科学基金项目(项目编号: LY24H090005)及浙江省重点研发计划项目(项目编号: 2025C01104)的资助。

参考文献

- [1] Yeo, K.K., Nagabushan, S., Dhall, G. and Abdelbaki, M.S. (2022) Primary Central Nervous System Germ Cell Tumors in Children and Young Adults: A Review of Controversies in Diagnostic and Treatment Approach. *Neoplasia*, **36**, Article 100860. <https://doi.org/10.1016/j.neo.2022.100860>
- [2] Liu, S., Ren, L., Gao, X., Hao, M. and Wang, G. (2022) Pathogenesis of Central Nervous System Germ Cell Tumors. *Frontiers in Oncology*, **12**, Article 991484. <https://doi.org/10.3389/fonc.2022.991484>
- [3] Malbari, F., Gershon, T.R., Garvin, J.H., Allen, J.C., Khakoo, Y., Levy, A.S., *et al.* (2016) Psychiatric Manifestations as Initial Presentation for Pediatric CNS Germ Cell Tumors, a Case Series. *Child's Nervous System*, **32**, 1359-1362. <https://doi.org/10.1007/s00381-016-3145-8>
- [4] Howell, M.D., Corrado, G.S. and DeSalvo, K.B. (2024) Three Epochs of Artificial Intelligence in Health Care. *Journal of the American Medical Association*, **331**, 242-244. <https://doi.org/10.1001/jama.2023.25057>
- [5] Theodosiou, A.A. and Read, R.C. (2023) Artificial Intelligence, Machine Learning and Deep Learning: Potential Resources for the Infection Clinician. *Journal of Infection*, **87**, 287-294. <https://doi.org/10.1016/j.jinf.2023.07.006>
- [6] Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., *et al.* (2019) Artificial Intelligence in Cancer Imaging: Clinical Challenges and Applications. *CA: A Cancer Journal for Clinicians*, **69**, 127-157. <https://doi.org/10.3322/caac.21552>
- [7] Currie, G. and Rohren, E. (2021) Intelligent Imaging in Nuclear Medicine: The Principles of Artificial Intelligence, Machine Learning and Deep Learning. *Seminars in Nuclear Medicine*, **51**, 102-111. <https://doi.org/10.1053/j.semnuclmed.2020.08.002>
- [8] Cruz-Gonzalez, P., He, A.W.J., Lam, E.P., Ng, I.M.C., Li, M.W., Hou, R., *et al.* (2025) Artificial Intelligence in Mental Health Care: A Systematic Review of Diagnosis, Monitoring, and Intervention Applications. *Psychological Medicine*, **55**, e18. <https://doi.org/10.1017/s0033291724003295>
- [9] Lüscher, T.F., Wenzl, F.A., D'Ascenzo, F., Friedman, P.A. and Antoniades, C. (2024) Artificial Intelligence in Cardiovascular Medicine: Clinical Applications. *European Heart Journal*, **45**, 4291-4304. <https://doi.org/10.1093/eurheartj/ehae465>
- [10] Schaye, V., Miller, L., Kudlowitz, D., Chun, J., Burk-Rafel, J., Cocks, P., *et al.* (2022) Development of a Clinical Reasoning Documentation Assessment Tool for Resident and Fellow Admission Notes: A Shared Mental Model for Feedback. *Journal of General Internal Medicine*, **37**, 507-512. <https://doi.org/10.1007/s11606-021-06805-6>
- [11] Lambeck, J., Hieber, M., Dreßing, A. and Niesen, W.D. (2019) Central Pontine Myelinosis and Osmotic Demyelination Syndrome. *Deutsches Ärzteblatt International*, **116**, 600-606. <https://doi.org/10.3238/arztebl.2019.0600>
- [12] Onem, Y., Terekeci, H., Kucukardali, Y., Sahan, B., Solmazgöl, E., Şenol, M.G., *et al.* (2010) Albumin, Hemoglobin, Body Mass Index, Cognitive and Functional Performance in Elderly Persons Living in Nursing Homes. *Archives of Gerontology and Geriatrics*, **50**, 56-59. <https://doi.org/10.1016/j.archger.2009.01.010>
- [13] Lee, S., Min, J., Kim, B., Ha, S.W., Han, J.H. and Min, K. (2021) Serum Sodium in Relation to Various Domains of Cognitive Function in the Elderly US Population. *BMC Geriatrics*, **21**, Article No. 328. <https://doi.org/10.1186/s12877-021-02260-4>
- [14] Aggleton, J.P., Nelson, A.J.D. and O'Mara, S.M. (2022) Time to Retire the Serial Papez Circuit: Implications for Space, Memory, and Attention. *Neuroscience & Biobehavioral Reviews*, **140**, Article 104813. <https://doi.org/10.1016/j.neubiorev.2022.104813>
- [15] Kupers, R.C., Fortin, A., Astrup, J., Gjedde, A. and Ptito, M. (2004) Recovery of Anterograde Amnesia in a Case of Craniopharyngioma. *Archives of Neurology*, **61**, 1948-1952. <https://doi.org/10.1001/archneur.61.12.1948>
- [16] Oertel, H., Schneider, H.J., Stalla, G.K., Holsboer, F. and Zihl, J. (2004) The Effect of Growth Hormone Substitution on Cognitive Performance in Adult Patients with Hypopituitarism. *Psychoneuroendocrinology*, **29**, 839-850. [https://doi.org/10.1016/s0306-4530\(03\)00151-3](https://doi.org/10.1016/s0306-4530(03)00151-3)
- [17] Nakamura, H., Takami, H., Yanagisawa, T., Kumabe, T., Fujimaki, T., Arakawa, Y., *et al.* (2022) The Japan Society for Neuro-Oncology Guideline on the Diagnosis and Treatment of Central Nervous System Germ Cell Tumors. *Neuro-*

- Oncology*, **24**, 503-515. <https://doi.org/10.1093/neuonc/noab242>
- [18] Acharya, S., DeWees, T., Shinohara, E.T. and Perkins, S.M. (2015) Long-Term Outcomes and Late Effects for Childhood and Young Adulthood Intracranial Germinomas. *Neuro-Oncology*, **17**, 741-746. <https://doi.org/10.1093/neuonc/nou311>
- [19] Eguia, H., Sánchez-Bocanegra, C.L., Vinciarelli, F., Alvarez-Lopez, F. and Saigí-Rubió, F. (2024) Clinical Decision Support and Natural Language Processing in Medicine: Systematic Literature Review. *Journal of Medical Internet Research*, **26**, e55315. <https://doi.org/10.2196/55315>
- [20] Parekh, A.D.E., Shaikh, O.A., Simran, Manan, S. and Hasibuzzaman, M.A. (2023) Artificial Intelligence (AI) in Personalized Medicine: AI-Generated Personalized Therapy Regimens Based on Genetic and Medical History: Short Communication. *Annals of Medicine & Surgery*, **85**, 5831-5833. <https://doi.org/10.1097/ms9.0000000000001320>
- [21] Brasil, S., Pascoal, C., Francisco, R., dos Reis Ferreira, V., A. Videira, P. and Valadão, G. (2019) Artificial Intelligence (AI) in Rare Diseases: Is the Future Brighter? *Genes*, **10**, Article 978. <https://doi.org/10.3390/genes10120978>
- [22] Alfaro, S., Liu, J., Naranjo Ortiz, C., Alfaro, A. and Lustberg, M. (2026) Applications of Machine Learning and Natural Language Processing to Neurocognitive Outcomes in Posttreatment Cancer Survivors: A Scoping Review. *Supportive Care in Cancer*, **34**, Article No. 71. <https://doi.org/10.1007/s00520-025-10293-3>
- [23] Vergun, S., Suhonen, J.I., Nair, V.A., Kuo, J.S., Baskaya, M.K., Garcia-Ramos, C., *et al.* (2018) Predicting Primary Outcomes of Brain Tumor Patients with Advanced Neuroimaging MRI Measures. *Interdisciplinary Neurosurgery*, **13**, 109-118. <https://doi.org/10.1016/j.inat.2018.04.013>
- [24] Hah, H. and Goldin, D.S. (2021) How Clinicians Perceive Artificial Intelligence-Assisted Technologies in Diagnostic Decision Making: Mixed Methods Approach. *Journal of Medical Internet Research*, **23**, e33540. <https://doi.org/10.2196/33540>
- [25] Monteith, S., Glenn, T., Geddes, J., Whybrow, P.C., Achtyes, E. and Bauer, M. (2022) Expectations for Artificial Intelligence (AI) in Psychiatry. *Current Psychiatry Reports*, **24**, 709-721. <https://doi.org/10.1007/s11920-022-01378-5>
- [26] Bedi, S., Liu, Y., Orr-Ewing, L., Dash, D., Koyejo, S., Callahan, A., *et al.* (2025) Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *Journal of the American Medical Association*, **333**, 319-328. <https://doi.org/10.1001/jama.2024.21700>
- [27] Savage, T., Wang, J., Gallo, R., Boukil, A., Patel, V., Safavi-Naini, S.A.A., *et al.* (2025) Large Language Model Uncertainty Proxies: Discrimination and Calibration for Medical Diagnosis and Treatment. *Journal of the American Medical Informatics Association*, **32**, 139-149. <https://doi.org/10.1093/jamia/ocae254>
- [28] Kolding, S., Lundin, R.M., Hansen, L. and Østergaard, S.D. (2024) Use of Generative Artificial Intelligence (AI) in Psychiatry and Mental Health Care: A Systematic Review. *Acta Neuropsychiatrica*, **37**, e37. <https://doi.org/10.1017/neu.2024.50>

附 录

Table S1. Descriptive statistics of R-IDEA scale for different models
表 S1. 不同模型 R-IDEA 量表描述性统计

模型类型	R-IDEA 量表	精神科门诊数据			精神科住院数据			多学科会诊住院数据		
		均值	中位数 (Q1, Q3)	众数	均值	中位数 (Q1, Q3)	众数	均值	中位数 (Q1, Q3)	众数
Grok 3	I 维度	3.25	3 (3, 4)	3.00	1.60	1 (1, 2.25)	1.00	1.85	1.5 (1, 3)	1.00
	D 维度	1.50	1.5 (1, 2)	2.00	1.45	1 (1, 2)	1.00	1.55	2 (1, 2)	2.00
	E 维度	1.80	2 (2, 2)	2.00	1.55	2 (1, 2)	2.00	1.45	1.5 (1, 2)	2.00
	A 维度	0.00	0 (0, 0)	0.00	1.35	1 (1, 2)	1.00	1.30	1 (1, 2)	1.00
	总分	6.55	7 (6, 7)	7.00	5.95	5.5 (5, 7)	5.00	6.15	6.5 (4, 8)	4.00
DeepSeek-R1	I 维度	3.30	3.5 (3, 4)	4.00	1.90	2 (1, 3)	1.00	1.85	1.5 (1, 3)	1.00
	D 维度	1.25	1 (1, 1.25)	1.00	0.15	0 (0, 0)	0.00	0.10	0 (0, 0)	0.00
	E 维度	1.75	2 (1.75, 2)	2.00	1.60	2 (1, 2)	2.00	1.80	2 (2, 2)	2.00
	A 维度	1.65	2 (1, 2)	2.00	1.75	2 (1.75, 2)	2.00	1.60	2 (1, 2)	2.00
	总分	7.95	8 (7, 9)	9.00	5.40	5 (4, 7)	4.00	5.35	5 (4.75, 6.25)	5.00
Gemini 2.0	I 维度	3.70	4 (3, 4)	4.00	2.05	2 (1, 3)	1.00	1.95	1.5 (1, 3)	1.00
	D 维度	1.75	2 (1.75, 2)	2.00	1.30	1 (1, 2)	1.00	1.90	2 (2, 2)	2.00
	E 维度	1.35	1 (1, 2)	1.00	1.45	1 (1, 2)	1.00	1.85	2 (2, 2)	2.00
	A 维度	1.25	1 (1, 1.25)	1.00	1.40	1 (1, 2)	1.00	1.50	1.5 (1, 2)	1.00
	总分	8.05	8 (7, 8)	8.00	6.20	5.5 (5, 8)	5.00	7.20	7 (6, 8.25)	6.00
ChatGPT-o3	I 维度	3.15	3 (2, 4)	4.00	1.60	1 (1, 2)	1.00	1.70	1 (1, 2)	1.00
	D 维度	1.25	1 (1, 1.25)	1.00	1.15	1 (1, 1.25)	1.00	0.15	0 (0, 0)	0.00
	E 维度	1.65	2 (1, 2)	2.00	1.40	1 (1, 2)	1.00	1.50	1.5 (1, 2)	1.00
	A 维度	1.45	1.5 (1, 2)	2.00	1.45	1 (1, 2)	1.00	1.20	1 (1, 2)	1.00
	总分	7.50	8 (6.75, 9)	8.00	5.60	6 (4.75, 6)	6.00	4.55	4.5 (3, 5.25)	3.00
Claude 3.7	I 维度	3.40	4 (3, 4)	4.00	1.75	1 (1, 2.25)	1.00	1.70	1 (1, 2.25)	1.00
	D 维度	1.40	1 (1, 2)	1.00	1.40	1 (1, 2)	1.00	1.40	1 (1, 2)	1.00
	E 维度	1.55	2 (1, 2)	2.00	1.95	2 (2, 2)	2.00	1.95	2 (2, 2)	2.00
	A 维度	1.05	1 (1, 1)	1.00	1.70	2 (1, 2)	2.00	1.75	2 (1.75, 2)	2.00
	总分	7.40	8 (7, 8)	8.00	6.80	7 (6, 7.25)	6.00	6.80	7 (6, 7.25)	6.00

注：Grok 3 为 Grok 第三代模型；DeepSeek-R1 为 DeepSeek 的 R1 模型；Gemini 2.0 为 Gemini 2.0 Flash Thinking 模型；ChatGPT-o3 为 ChatGPT 的 o3 模型；Claude 3.7 为 Claude 3.7 Sonnet 模型。