基于大语言模型的中国民航公务员考试测评

杨凯杰[1],秦雪峰[2a],莫济懋[2b],王楚为[3],李冠霖[2a],*DING H.Q. Chris*[4],蔡元哲[2a]
(1.利物浦大学,利物浦,L697ZX; 2.深圳技术大学 a.人工智能学院; b.工程物理学院,广东 深圳 518118;
3.厦门大学马来西亚分校,吉隆坡 43900; 4.香港中文大学(深圳),广东 深圳 518172)

摘 要: 本文对包括ChatGPT-4o在内的5种大型语言模型(LLM,large language model)在中国民航国家公务员考试(NCSE,national civil servant exam)中的应试能力进行了系统评估与分析。研究选取2022—2024年NCSE真题,以预设的标准化提问范式向5种LLM分别输入题目并记录其输出结果,进而统计5种LLM的答题正确率以衡量其综合能力。实验结果显示,DeepSeek-V3、DeepSeek-R1、ChatGPT-4o、Gemini-1.5 Flash、ERNIE Bot-4.0 Turbo总分分别为145.20、145.41、127.47、107.56、86.40,除ERNIE Bot-4.0 Turbo之外,均高于人类考生平均成绩93.50,其中DeepSeek-V3、DeepSeek-R1的分数达到NCSE的高分区间。此外,本文深入讨论5种LLM的优势与不足,对常识判断、言语理解与表达、数量关系、判断推理、资料分析等不同题型的答题表现进行了细分对比,并归纳了5种LLM在应对复杂逻辑推理与多步骤运算题目时的典型错误类型。

关键词: 大语言模型(LLM);国家公务员考试(NCSE);模型性能评估
中图分类号: TP18 **文献标志码:** A **文章编号:** 1674-5590(2025)06-0088-09

Chinese aviation national civil servant exam assessment using large language model

YANG Kaijie[1], *QIN Xuefeng*[2a], *MO Jimao*[2b], *WANG Chuwei*[3],
LI Guanlin[2a], *DING H.Q. Chris*[4], *CAI Yuanzhe*[2a]
(1. *University of Liverpool, Liverpool* L697*ZX*, *UK*; 2a. *College of Artificial Intelligence*; 2b. *College of Engineering Physics*, *Shenzhen University of Technology*, *Shenzhen* 518118, Guangdong, *China*; 3. *Xiamen University Malaysia*, *Kuala Lumpur* 43900, *Malaysia*;
4. *The Chinese University of Hong Kong* (*Shenzhen*), *Shenzhen* 518172, Guangdong, *China*)

Abstract: This paper systematically evaluates and analyzes the performance of five large language models (LLMs), including ChatGPT-4o on the Chinese national civil servant exam (NCSE). The study selected NCSE past exam papers from 2022 to 2024, input the questions into the five LLMs using a predefined standardized query format, recorded their outputs, and calculated their accuracy rates to assess their overall capabilities. The experimental results show that the total scores for DeepSeek-V3, DeepSeek-R1, ChatGPT-4o, Gemini-1.5 Flash, and ERNIE Bot-4.0 Turbo are 145.20, 145.41, 127.47, 107.56, and 86.40, respectively, except for ERNIE Bot-4.0 Turbo, all models scored higher than the average human candidate score of 93.50. Among them, DeepSeek-V3 and DeepSeek-R1 achieved scores within the high-score range of NCSE. Furthermore, this paper delves into the strengths and weaknesses of the five LLMs, provides a detailed comparison of their performance across different question types, such as common sense judgment, verbal comprehension and expression, quantitative relationships, judgment and reasoning, and data analysis, and summarizes typical error patterns when handling complex logical reasoning and multi-step calculation problems.

Key words: large language model (LLM); national civil servant exam (NCSE); model performance evaluation

大型语言模型(LLM,large language model)正以前所未有的速度发展,其在理解、生成和推理方面的卓越能力,不仅重塑了自然语言处理(NLP,natural language processing)领域,更被视为迈向通用人工智能(AGI,artificial general intelligence)的关键一步。LLM已应用至民航、教育、医疗等多个行业,展现出解决

收稿日期:2025-09-10; **修回日期:**2025-11-20
基金项目:深圳市高等院校稳定支持计划项目(20231127194506001);广东省高校创新项目(2024KTSCX055)
作者简介:杨凯杰(2004—),男,河北唐山人,本科生,研究方向为人工智能、大模型评价.

复杂现实问题的巨大潜力，并逐渐成为大众获取信息、辅助决策的重要工具之一。与此同时，LLM 在模拟人类认知能力方面的表现，特别是在标准化考试中的应用，已成为学术界关注的热点。已有研究表明，部分 LLM 在美国执业医师资格考试（USMLE，United States medical licensing examination）等专业测试中取得了令人瞩目的成绩[1]。

中国民航公务员是负责民用航空事业管理与监管的国家工作人员，其承担着行业规划、安全监管、政策制定、市场监督、适航审定及国际合作等关键职责，是确保民航安全高效运行、推动行业高质量发展和国家航空管理体系的重要基石。中国民用航空局通过中国国家公务员考试（NCSE，national civil servant exam）公开招聘选拔中国民航公务员。

在此背景下，本文旨在系统评估以 DeepSeek-R1为代表的主流 LLM 在 2022—2024 年中国 NCSE 中的表现。本文不仅关注主流 LLM 能否达到“通过”标准，更致力于深入分析其在各项测试模块上的具体表现差异，探讨模型架构与训练数据对处理此类高度情境化、知识密集型任务的影响。本文的研究核心目标是通过研究 LLM 在中国民航公务员考试中的表现，来揭示 LLM 在应对复杂、本土化、标准化测试时的优势与局限。

1　研究背景

LLM 是能够理解和生成人类语言的计算模型。LLM 的飞速发展，已显著提升了 NLP 的能力边界，推动了从传统统计方法到深度神经网络模型的范式转变[2-3]。LLM 不仅在文本理解与生成方面表现出色，更展现出解决复杂、知识密集型任务的潜力，促使研究者开始探索其在模拟人类智能，特别是应对标准化考试方面的能力。评估不同 LLM 在各种考试中的表现，已成为衡量其认知水平、泛化能力和实际应用价值的关键途径。

近年来，研究学者们开始将 LLM 应用于各类标准化考试的评估中，考查其面对标准化考题的解题能力。现有研究已初步证实，顶尖的 LLM 在多种国际高难度专业考试中取得了令人欣喜的成绩。例如，Katz 等[4]证实 ChatGPT-4o 能够通过美国律师资格考试；Skalidis 等[5]的研究表明 ChatGPT-4o 在欧洲核心心脏病学考试中获得了及格成绩；Tsoutsanis 等[6]的研究显示，Bing Chat 在多学科招聘评估测试中的表现超过了人类考生的平均水平。与此同时，针对中文语境下的专业资格测试研究，Xu 等[7]评估了国产大模型在中国护士执业资格考试中的表现，发现特定模型已具备通过考试的能力；Hong 等[8]则提出了针对中国各类职业资格认证的综合评测基准，进一步验证了 LLM 在国内垂直领域的应用潜力。上述国内外研究充分展示了 LLM 在处理高难度专业语言任务中的能力。

NCSE 以其独特的复杂性以及对精微语言理解和文化背景知识的要求而著称，加之不同 LLM 在训练策略、模型架构、解码策略上的固有差异导致其能力表现各异，使其在 NCSE 上的适用性成为一个亟待验证的问题[9]。此外，不同 LLM 在“有用性”与“安全性”之间的权衡策略差异显著。通过人类反馈强化学习（RLHF，reinforcement learning from human feedback）对齐的 LLM 在面对敏感指令时可能会表现出过度防御（over-refusal）的倾向[10]，这一特性在处理公务员考题时尤为关键。尽管 LLM 在部分标准化测试中已展现潜力[4-6]，且现有研究推出了如 C-Eval[11]和 CMMLU[12]等针对中文语境的综合性评估基准，但这些基准更侧重于考察中文人文社科及理工科的基础知识广度，未能深入覆盖公务员考试中特有的行政职业能力逻辑与申论写作的具体要求。因此，系统比较不同 LLM 在 NCSF 这一复杂场景的表现尚属空白。本文核心旨在评估当前代表性的 LLM，如DeepSeek-R1 能否应对NCSE 的综合要求，并重点对比分析这些具有不同特性的模型之间的能力差异，以期为理解LLM 在处理高难度、本土化及知识密集型任务时的真实界限与潜力提供必要的实证参考。

2　研究方法

2.1　数据集

本研究选取了 2022—2024 年 NCSE 的完整试题作为评估材料，通过人工表格统计测试题目，将其中的重复数据进行排除。其中，行政职业能力测验（简称行测）部分共包含 130 道单项选择题，按能力模块划分为 5 个子部分：常识判断、言语理解与表达、数量关系、判断推理及资料分析，需要考生在 2 h 内完成。

行测部分题型覆盖内容及考查方向具体如下：①常识判断涵盖人文、科技、法律、历史等多学科领域，考查考生的基础知识储备与综合素养；②言语理解与表达测试语言信息的理解能力，包括词义辨析、句意分析、段落逻辑等；③数量关系涉及基础数学运算与数字推理能力，要求考生具备一定的逻辑计算能力；

④判断推理主要考查逻辑思维与演绎推理能力，题型包括图形推理、定义判断、类比推理等；⑤资料分析侧重数据提取、计算与综合分析能力，是评估信息处理能力的重要部分。行测试卷组成如表1 所示。

表 1　中国 NCSE 试卷组成

Tab.1　Composition of the NCSE paper of China

题型构成	题目数量	分值/题	总分	答题时间/h
行测–常识判断	20	0.50	10	2
行测–言语理解与表达	40	0.08	32	
行测–数量关系	15	0.08	12	
行测–判断推理	40	0.65	26	
行测–资料分析	20	1.00	20	
申论	5	10/15/20/20/35	100	3

申论部分则采用材料写作形式，由 5 则材料及对应的 5 道主观题组成，主要考查考生对给定材料的信息提取、概括、梳理、归纳与提炼能力，并要求考生在 3 h 内基于材料提出观点，并最终形成结构完整、论证充分的议论文，申论试卷组成如表 1 所示。

2.2　参评模型

本文旨在基于 NCSE 对当前具有较高影响力及先进性的 LLM 进行评估。LLM 的选取主要遵循以下原则：①在学术界与产业界具有广泛的关注度和应用基础；②在公开的评估基准或相关研究中展现出领先或具有代表性的性能[13]；③具有公开可及性或相关技术报告的可用性，允许进行有效的分析与比较[14–18]。除上述因素外，本研究特别考虑了国内 LLM，因为国内 LLM 通常会使用大量中文数据集进行训练，而海外 LLM 则可能会选取全球范围内的数据资源[19]。

基于上述原则，本研究最终确定纳入评估的 LLM 共计 5 种，包括 ChatGPT–4o[14]、Gemini–1.5 Flash[18]（Gemini）、ERNIE Bot–4.0 Turbo[17]（文心一言）、DeepSeek–V3[15]与 DeepSeek–R1[16]。参评模型基本信息如表 2 所示。

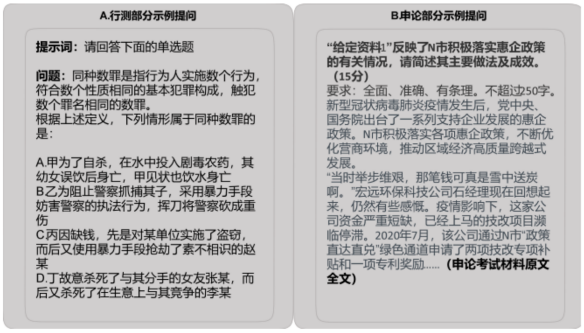
表 2　参评模型信息

Tab.2　Information on participating models for evaluation

模型名称	开发机构	发布时间	模型特点
ChatGPT–4o	OpenAI	2024.05	多模态大模型，支持输入输出文字/图片/文件
Gemini–1.5 Flash	Google	2024.09	记忆力强（支持超长对话），响应快，性价比高
ERNIE Bot–4.0 Turbo	百度	2024.06	记忆力强（支持超长对话），支持图像、文件等多模态信息处理
DeepSeek–V3	深度求索	2024.12	采用混合专家（MOE）架构，实现高效推理，经济训练
DeepSeek–R1	深度求索	2025.01	引入大规模强化学习（RL）训练，实现思维链，展现出推理能力

2.3　提问范式

既有研究已充分证明，提示词（prompt）的设计对 LLM 的输出质量和行为具有显著影响[20]。为确保评估结果的稳定性，减少因提示词差异引入的实验偏差，本文对所有评测实例采用了统一的输入范式。具体而言，每个输入范式均遵循“提示词+原题目+原题选项”的结构化格式。此标准化格式的详细结构与示例，如图 1 所示。



（a）行测　　（b）申论

图 1　提问范式

Fig.1　Prompting paradigm

2.4　评估步骤

本文采用标准化输入流程，将规范化试题以单题为单位逐次输入 LLM 的网页端并对答题结果进行评分。为控制混杂变量并消除上下文信息对生成结果的干扰，实验严格遵循以下操作规程：每次仅输入单道题目，待 LLM 完成响应后立即保存输出记录；随后启动独立浏览实例（新标签页/新会话），彻底清除上下文缓存后再加载下一题项。该方法可降低情境线索对 LLM 性能评估的潜在偏倚风险，从而增强了实验结果的可验证性与跨场景适用性。

2.5　数据分析

在获得 5 种 LLM 对全部试题的作答结果后，课题组对输出结果开展了系统性评分。评分标准严格参照官方发布的参考答案制定，通过逐项比对模型生成答案与标准答案的一致性程度，确定每道题目的答题准确性。为解析模型在多维认知领域的性能特征，本文基于原始试卷题型分类，对试题库实施分项评估。具体而言，将试题划分为常识判断、言语理解与表达、数量关系、判断推理及资料分析 5 大模块，分别统计 5 种 LLM 在不同模块的得分分布特征。该分析框架不仅能够量化 LLM 的能力异质性特征，还可直观反映其在标准化考试场景中的优势与局限。在数据统计方面，本文统计了 5 种 LLM 在各类题型中的正确率，其数学表达式表示为

$$r = \frac{\alpha - \beta}{\alpha} \times 100\% \quad (1)$$

式中：r 为模型正确率；α 为答题总数；β 为错误回答题目数。

2.6 申论专家评分标准

为确保对 LLM 生成的申论内容进行专业且客观的评估，课题组特邀相关专家承担评分工作。这些专家在马克思主义理论、公共策论等领域具备深厚学术背景与丰富的评审经验。评分严格参照既有的申论评价规范，从政治立场、价值导向、理论运用、政策把握等方面进行考虑，对 LLM 生成内容的思想观点、论证结构、语言表达及创新性进行综合评判，力求评分结果的权威性与准确性。

3 结果与分析

3.1 总体考试成绩分析

5 种 LLM 在 2022—2024 年的 NCSE 平均得分，如表 3 所示。

表 3 不同模型在 NCSE 的平均得分（2022—2024）

Tab.3 Average score of different models in NCSE (2022—2024)

模型名称	行测得分	申论得分	总得分
ChatGPT-4o	66.30	61.17	127.47
Gemini	49.72	57.84	107.56
文心一言	60.90	25.50	86.40
DeepSeek-V3	81.93	63.27	145.20
DeepSeek-R1	85.25	60.16	145.41

从表 3 可以看到，DeepSeek 系列模型取得了较好的分数，除了文心一言以外，其他 LLM 的 NCSE 平均得分远高于人类考生的平均分 93.50 分。对于NCSE来说，其分数总体分布呈“正态分布”（钟形曲线）特点：高分区间（135 分以上）占比很小；中等偏上区间（120~135 分）是进入面试分数线的核心竞争区；中等区间（100~120 分）是考生人数最为集中的区域；对于许多竞争比例适中或偏冷的岗位，面试分数线可能在此区间；低分区间（100 分以下）人数也较多，但通常低于官方公布的合格线（行测单科也有合格线），除非是竞争极小的偏远冷门岗位或特殊专业岗位，否则很难进入面试分数线。其中，DeepSeek-V3（145.20分）和DeepSeek-R1（145.41 分）都达到了 NCSE 的高分区间，一般情况下都可以进入面试。ChatGPT-4o（127.47 分）达到中等偏上的考生水平，可以进入面试分数线的核心竞争区，例如一些一线城市的冷门岗位或者二线城市的热门岗位。Gemini（107.56 分）只达到中等区间，只能达到一些二三线城市冷门岗位的面试分数线。文心一言的成绩基本上无法进入面试环节。

3.2 行测结果分析

3.2.1 常识判断

为了多维度评估 LLM 在常识判断任务中的能力，本文使用“粉笔”网站试题分类，对 5 种 LLM 在各题型分类中的正确率进行了统计，常识判断题型测试结果如表 4 所示。常识判断中 ChatGPT-4o、Gemini、文心一言、DeepSeek-V3 和 DeepSeek-R1 的平均正确率分别是 78.17%、77.78%、92.06%、95.92%和 97.78%。需要指出的是，在测试题目中由于涉及政治敏感等因素，DeepSeek-V3 和 Gemini 各有 11 道题没有回答，DeepSeek-R1 有 15 道题没有回答，因此，本文在统计上述最终正确率时剔除了这部分题目。

表 4 常识判断题型正确率（2022—2024）

Tab.4 Accuracy rate of common sense judgment question types (2022—2024)

%

题型分类	ChatGPT-4o	Gemini	文心一言	DeepSeek-V3	DeepSeek-R1	人类考生
政治常识	100.00	66.67	100.00	88.89	100.00	52.68
经济常识	100.00	100.00	100.00	100.00	100.00	34.00
科技常识	33.33	33.33	66.66	100.00	100.00	41.53
人文常识	50.00	0	100.00	100.00	100.00	47.45
地理常识	100.00	100.00	100.00	100.00	100.00	38.40
法律常识	85.71	90.00	85.71	95.00	90.00	38.89

在所测试的 3 年行测题目中，ChatGPT-4o 共获得 78.17%的平均正确率，整体表现优异，尤其在政治常识类题目中取得了 100%的平均正确率，显示其在处理政治常识相关知识方面具有较强的理解能力。Gemini 的平均正确率为 77.78%，表现较为一般，刚刚到达及格表现。文心一言获得较好的平均正确率，为 92.06%，表现出广泛的知识覆盖能力与较强的综合理解能力。DeepSeek-V3 模型和 DeepSeek-R1 模型则表现较为良好，都有着超过 95%的平均正确率。DeepSeek 模型在常识推断任务中的卓越表现，可能是由于其训练数据更多基于中文语料库且模型架构更加先进。

综合 5 种 LLM 的得分情况可见，不同 LLM 在通识知识处理能力方面存在显著差异。DeepSeek 的 2 个模型总体表现最为优异，几乎在所有知识模块中均展现出高水平的理解和答题能力；文心一言紧随其后，在多领域表现出色，具有较强的泛化能力；Gemini 表现稍逊于 ChatGPT-4o。

除此之外，研究发现 DeepSeek 模型的不同版本间存在性能差异。在常识类题型问答评估中，DeepSeek-

R1 模型较 DeepSeek-V3 模型正确率更高，但DeepSeek-R1 更容易出现拒绝回答题目的现象，DeepSeek-V3 版本则回答了更多的题目。初步分析表明，DeepSeek-R1 在处理特定查询时可能更容易触发内容安全审查机制，导致响应中止。此现象或与其采用的思维链（chain-of-thought）推理过程有关，该过程在生成中间步骤时可能增加了触及敏感边界的概率，导致作答中断，放弃作答。因此，在常识类任务中观察到的性能差异，需审慎解读，其可能并非完全反映 LLM 的基础推理能力。

常识判断部分的测试结果表明：在记忆性知识的精确掌握方面，人类表现与 LLM 之间存在显著差距。鉴于 LLM 在此类任务中已展现出高度的准确性与稳定性，对于公务员在日常工作中遇到的此类信息查询与验证需求，可以考虑引入 LLM 作为智能化辅助手段，这不仅能提升工作效率，亦是确保信息严谨性的重要途径。

3.2.2　言语理解与表达

言语理解与表达测试结果如表 5 所示，表 5 中列出了 2022—2024 年 5 种 LLM 在“言语理解与表达”模块下各子题型中的平均正确率。5 种 LLM 在该模块展现了不同程度的语言理解与逻辑分析能力，多数模型表现突出。

表 5　言语理解与表达题型平均正确率（2022—2024）

Tab.5　Accuracy rate of verbal comprehension and expression question types (2022—2024)

%

题型分类（对应题数）	ChatGPT-4o	Gemini	DeepSeek-R1	DeepSeek-V3	文心一言
逻辑填空	68.35	51.65	88.35	86.65	58.35
片段阅读	69.73	69.73	90.91	87.91	87.91
语句表达	88.89	74.11	85.22	88.89	88.89

其中，DeepSeek-R1 与 DeepSeek-V3 的平均正确率分别为 88.16%与 87.82%，在 5 种 LLM 中排名靠前，表明其在处理中文言语理解任务方面具有显著优势，能够较为准确地完成与词句辨析、段落推理等相关的问题。其余 3 种 LLM 平均的正确率从高到低依次为：文心一言（78.38%）、ChatGPT-4o（75.65%）、Gemini（65.16%）。除 Gemini 以外，其余 4 种 LLM 的总体表现均超越了人类考生的平均正确率（69.20%），这凸显了 LLM 在语义分析与语言逻辑推理方面的坚实能力。

总体而言，LLM 在处理“言语理解与表达”类题目时展现出一定的语义理解能力。然而，不同模型在该模块中的表现仍存在明显差异，其核心差异在于是否具备有效的语境分析能力。以 Gemini 为例，其在答题过程中常缺乏对上下文语义环境的准确把握，无法识别题目中关键语言线索，导致理解偏差，反映出其语义解析能力较弱。相较之下，2 个 DeepSeek 模型在答题时表现出明显的语境推理与语言整合过程，能够将词语含义结合上下文进行综合判断，遵循较为完整的语义推理路径，体现出较强的语义理解逻辑能力。对于 LLM 的不同表现，本文得出了以下 3 点原因：训练语料的规模；训练方向的差异；训练强度与策略差异。综上，LLM 在“言语理解与表达”模块上的性能差异反映出其在语义建模深度、语言上下文整合能力以及语言逻辑推理机制方面的显著差异，这些差异本质上可归因于其训练数据与架构策略的不同。

3.2.3　数量关系

表 6 展示了 5 种 LLM 在 2022—2024 年的数量关系题型下的答题表现。与人类考生在面对不同类型数学题目时表现出能力差异相似，LLM 亦呈现出显著的性能波动。

表 6　数量关系题型正确率（2022—2004）

Tab.6　Accuracy rate of quantitative reasoning types (2022—2024)

%

年份	ChatGPT-4o	Gemini	文心一言	DeepSeek-V3	DeepSeek-R1
2022	50.00	20.00	40.00	80.00	80.00
2023	60.00	40.00	30.00	60.00	70.00
2024	70.00	30.00	50.00	50.00	70.00

从结果来看，DeepSeek-R1 整体表现最为优异，在 2022—2024 年数量关系题型均取得最高正确率。这一优势可能源于其在数学计算与逻辑推理能力方面的建模更为精细。此外，通过对错误答案的分析发现，多数 LLM 在解题过程中出现了题意理解偏差，而 DeepSeek-R1 在这方面表现出更强的上下文理解与意图捕捉能力，可能是其答题准确率较高的关键因素。

整体而言，虽然 LLM 在数学类推理任务中表现出显著的进步潜力，且呈现出随时间优化的趋势，但其在面对结构复杂、语义歧义或多步计算要求的题目时，仍面临诸多挑战。进一步提升 LLM 的数学建模能力、上下文解析能力及推理链条的完整性，是未来研究与优化的关键方向。

3.2.4　判断推理

在公务员考试判断推理模块的各个子任务中，5 种 LLM 分别有如图 2 的平均正确率。不同 LLM 展现出显著的能力差异。

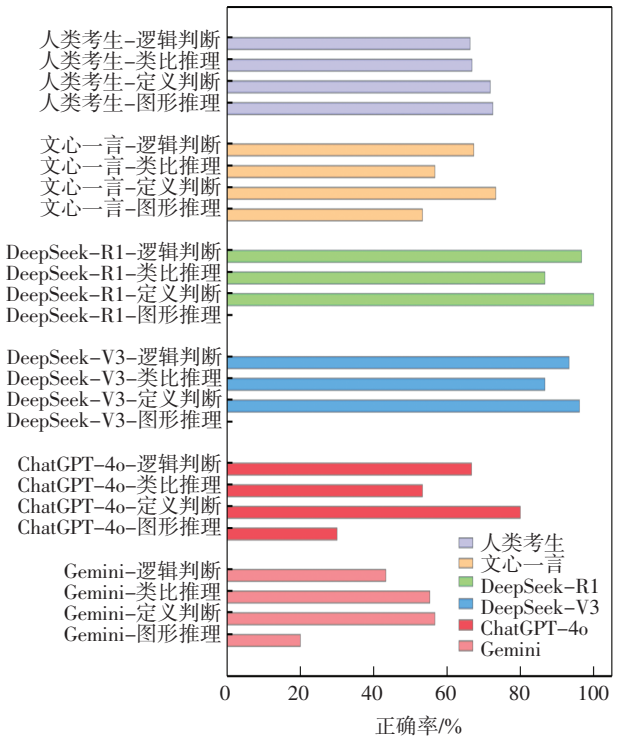


图 2　判断推理题型正确率

Fig.2　Accuracy rate of judgment and reasoning types

DeepSeek 系列模型（R1 与 V3 版本）在此次评估中表现最优。DeepSeek-R1 在定义判断、类比推理和逻辑判断任务中的正确率分别达到 100.00%、86.70%和 96.70%；DeepSeek-V3 在 3 项任务中的正确率依次分别为 96.15%、86.70%和 93.30%。

文心一言以 62.50%的平均正确率排名第三，其在图形推理和类比推理子任务中的正确率分别为 53.30%和 56.70%。该模型在定义判断与逻辑判断任务中分别取得 73.30%和 67.30%的正确率。相较于其他 LLM，文心一言的各项能力指标更为均衡（方差相对较低），显示出较稳定的综合性能。

ChatGPT-4o 以 57.50%的平均正确率位列第四，在定义判断任务中表现出色（80.00%正确率），但在图形推理任务中仅获 30.0%正确率，成为其整体表现的主要短板。Gemini 模型表现相对较弱，4 项任务总平均正确率为 43.30%，其中图形推理任务得分最低（20.00%），定义判断任务得分最高（56.70%）。

值得注意的是，人类考生在该模块的平均正确率为 69.35%（其中图形推理 72.50%、逻辑判断66.30%）。DeepSeek-R1（70.85%）的表现超出人类考生基准线，其余 LLM 均低于人类考生水平。人类考生表现呈现显著均衡性特征：所有子任务正确率均稳定超过60.00%阈值，且个体间差异幅度较小。

研究显示，大部分先进的 LLM（尤其是 DeepSeek 系列）在 NCSE 判断推理模块已具备超越人类考生的文本推理能力，拥有较完整的逻辑推演能力，可以思考并解决逻辑问题。

3.2.5　资料分析

在“资料分析”模块的测试中，5 种 LLM 在处理涉及图表理解、数据信息提取与综合计算类题型时平均正确率如表 7 所示。具体而言，DeepSeek-R1 与 DeepSeek-V3 分别在该模块取得了 100.00%和93.33%的平均正确率；与之对比，具备多模态处理能力的 ChatGPT-4o、文心一言及 Gemini 的平均正确率分别为 61.67%、30.00%和26.67%。人类考生在该完整模块上的平均正确率为 72.90%。

表 7　资料分析题型平均正确率（2022—2024）

Tab.7　Average accuracy rate of data analysis types (2022—2024)

%

模型	2022 年	2023 年	2024 年	2022—2024 年
ChatGPT-4o	55.00	65.00	65.00	61.67
文心一言	35.00	20.00	35.00	30.00
Gemini	30.00	30.00	20.00	26.67
DeepSeek-V3	80.00	100.00	100.00	93.33
DeepSeek-R1	100.00	100.00	100.00	100.00

尽管 DeepSeek 模型在纯文本资料分析任务上展现了卓越性能，但其高分仅反映了在限定任务范围内的能力。而其他 LLM（尤其是 ChatGPT-4o）虽然整体得分较低，但分数代表了其在更全面，且包含多模态信息处理的数据分析任务上的综合表现。这一结果不仅揭示了不同 LLM 在数据分析能力上的差距，更凸显了多模态能力对于 LLM 完成真实世界复杂数据分析任务的重要性，当前直接比较不同能力范围 LLM 得分的做法需格外审慎。

3.2.6　实验小结

最后，根据上述数据，可得出 5 种 LLM 在2022—2024 行测试题中的平均正确率，Gemini 为 48.80%，DeepSeek-V3 为 81.13%，DeepSeek-R1 为 83.83%，文心一言为 66.89%，ChatGPT-4o 为 68.68%。其中DeepSeek-R1 得分最高。LLM 的答题成绩高于人类考生，说明 LLM 在行测考试方面是优于人类考生的。其主要原因在于行测中有大量的记忆过程，所以LLM 取得了不错的成绩。

3.3　申论

在申论的测试中，5 种 LLM 在申论考试中不同材料的具体得分如表 8 所示。研究重点选取了在该模块中表现最为突出的 DeepSeek-V3 模型进行深入分析。测试结果显示，DeepSeek-V3 在申论任务中取得了

63.27 总分，这一成绩不仅高于 DeepSeek-R1 的 60.16 分，并且显著超过了人类考生 30.38 总分的平均水平。尽管与人类考生中的最高得分 87 分相比尚有差距，并且除文心一言外，其余受测模型均超过了人类平均分，但 DeepSeek-V3 的表现在材料理解、信息整合与文字表达方面已展现出较强的能力。

表 8　申论题型平均得分（2022—2024）

Tab.8　Average scores of essay types (2022—2024)

题目	ChatGPT-4o	DeepSeek-R1	Gemini	DeepSeek-V3	文心一言
材料 1	5.33	5.83	4.00	6.77	0.33
材料 2	7.17	9.00	6.17	10.33	2.17
材料 3	10.17	10.50	10.17	11.50	4.50
材料 4	14.33	12.50	13.33	11.67	5.33
材料 5	24.17	22.33	24.17	23.00	13.17

尽管 DeepSeek-V3 取得了令人鼓舞的成绩，但其表现相较顶尖人类考生，仍显现出一定的不足。这些不足主要体现在政策理解的深刻程度、语言运用的灵活性及篇章结构的逻辑严谨性等方面。这表明，尽管 DeepSeek-V3 在处理文字综合任务和书面表达方面具备了较高的实际应用潜力，但在更为高阶的写作技巧、深层语境的精确把握以及复杂的价值判断等层面，仍有进一步提升的空间。

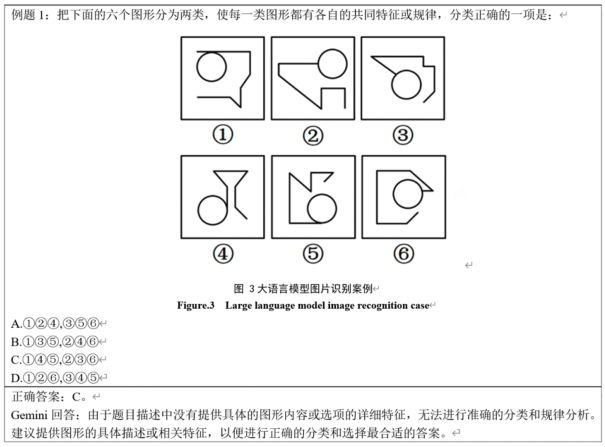
总体而言，DeepSeek-V3 在申论任务中的表现不仅反映了当前 LLM 在结构化写作和政策相关文本生成能力上的显著进步，同时也揭示了其作为高效写作辅助工具的广阔应用前景，特别是在诸如材料的初步梳理、核心观点的提炼与引导等工作环节，LLM 已能提供有力的支持。

3.4　结果比较与案例分析

在对 5 种 LLM 进行系统测试后，课题组对其答题行为和生成内容进行了观察与归纳，发现在此过程中，部分 LLM 展现出一些独特的作答特征及异常现象。由于这些现象难以通过纯粹的数据统计进行量化呈现，故在本节中将结合具体示例进行深入分析，以补充实验结果中未被量化体现的认知偏差与系统性误差。

3.4.1　LLM 会由于图片识别产生推理偏差

部分题型如涉及复杂图像识别或视觉信息深度提取的题目，显著超出了当前多数 LLM 的核心能力边界。在例题 1 中，如图 3 所示，题目需要考生找出图形之间的规律，并正确分类。参与测试的 LLM 普遍暴露出图像内容识别不准确、甚至完全错误的问题，进而导致其回答内容严重偏离题意。这一现象清晰地揭示了当前 LLM 在处理多模态信息，尤其是视觉理解与推理任务时的显著局限性。这种局限性具体表现为多个方面，例如，在图形算术或资料分析等题型中，LLM 常错误地识别图形中标注的关键数据（如将数字“6”误认为“8”或遗漏小数点），这直接导致后续的计算和推理过程出现重大偏差，最终得出与正确答案大相径庭的结果。


例题 1：把下面的六个图形分为两类，使每一类图形都有各自的共同特征或规律，分类正确的一项是：

①　②　③

④　⑤　⑥

A.①②④,③⑤⑥

B.①③⑤,②④⑥

C.①④⑤,②③⑥

D.①②⑥,③④⑤

正确答案：C。

Gemini 回答：由于题目描述中没有提供具体的图形内容或选项的详细特征，无法进行准确的分类和规律分析。建议提供图形的具体描述或相关特征，以便进行正确的分类和选择最合适的答案。

图 3 大语言模型图片识别案例

Figure.3　Large language model image recognition case

图 3　LLM 图片识别案例

Fig.3　Image recognition case of LLM

更进一步，对于一些并非源于真实世界物体的抽象几何图形或逻辑关系图，这些模型的理解障碍尤为突出。以图 3 展示的图形推理题为例，LLM 往往难以像人类一样洞察其中隐含的规律或逻辑模式，在部分情况下甚至会因无法有效解析图形而表现出“放弃”正常推理的倾向，或给出随机性、无逻辑性的答案，其表现远逊于能够通过观察和归纳找到解题思路的人类考生。同时，LLM 倾向于将试题中出现的抽象符号或不规则形状与其知识库内已有的现实世界具体物体进行不恰当的联想，从而产生识别和理解上的偏差。不仅如此，对于图像中诸如线段相对长短、图形面积相对大小、位置关系等需要依赖精细视觉感知和比较才能准确获取的信息，LLM 也时常表现出处理困难。这些对于人类考生而言通常较为直观和容易判断的视觉特征，却往往成为大语言模型准确理解题意的障碍。

这些在图像识别和视觉信息处理方面的普遍性不足，构成了当前 LLM 在应对包含视觉元素考题时的一个核心瓶颈。这不仅极大地影响了其在相关题目上的作答准确性，也明确指出了 LLM 在提升多模态理解与推理能力方面亟待加强的方向。

3.4.2　LLM 会拒绝回答政治敏感的问题

另一个值得关注的现象是，在面对涉及政治敏感

内容的题目时，部分模型会选择回避性作答、模糊表达，甚至完全跳过题目不予回答。

例如，图 4 中例题 2 要求考生回答中国政治决策的问题，其中包含国家政策改革方面的问题。DeepSeek–V3 与 DeepSeek–R1 均对 2022 年的申论考试的材料一做出了“无法回答”的响应，同时，Gemini 在行测的部分题型中也出现了这种现象。其原因可能是 LLM 判定材料中可能包含政治敏感内容。这种行为可能与模型训练阶段预设安全机制以及在网页端应用部署时的内容过滤策略有关，这些机制导致 LLM 在特定情况下无法做出实质性回答。值得一提的是，在本地部署的DeepSeek 模型可以作答该问题，这可能说明网页端应用的内容过滤策略更加严格，在处理相关问题时建议使用本地部署模型。

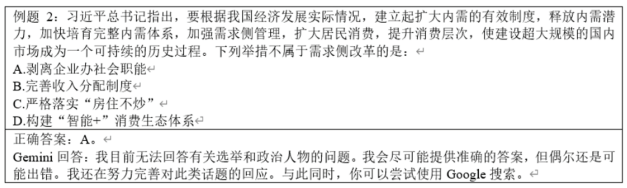

例题 2：习近平总书记指出，要根据我国经济发展实际情况，建立起扩大内需的有效制度，释放内需潜力，加快培育完整内需体系，加强需求侧管理，扩大居民消费，提升消费层次，使建设超大规模的国内市场成为一个可持续的历史过程。下列举措不属于需求侧改革的是：↵
A.剥离企业办社会职能↵
B.完善收入分配制度↵
C.严格落实“房住不炒”↵
D.构建“智能+”消费生态体系↵
正确答案：A。↵
Gemini 回答：我目前无法回答有关选举和政治人物的问题。我会尽可能提供准确的答案，但偶尔还是可能出错。我还在努力完善对此类话题的回应。与此同时，你可以尝试使用 Google 搜索。↵

图 4　政治常识题例题

Fig.4　Example of a political common sense question

尽管这些内容在实际考试环境中通常并不构成敏感话题，部分 LLM 仍因其内部策略而选择不做处理，导致其在多道题目中未能给出有效答案，或作答严重偏离题意，从而显著影响了其整体测试表现。此类现象不仅反映了当前模型在特定类型题目上的答题能力局限，也为未来 LLM 的优化方向提供了重要参考依据。如何在有效保障生成内容安全性的同时，提升 LLM 在特定文化语境与国家治理相关知识背景下的适应能力与作答灵活度，将成为后续模型设计中的关键问题。

3.4.3　LLM 的文字生成功能容易超过要求字数

除前述的推理与作答不一致的现象外，部分 LLM 在处理某些任务时，其生成内容的方式也表现出一些特点。在主观的写作任务，如申论部分的测试中，其会出现忽视格式要求的现象，如超出字数等问题。

在申论模块测试中，研究人员为模型输入了统一的情景提要与写作要求（其中包括了字数限制）。选取 DeepSeek 的 2 个模型进行深入分析，结果显示，2 个模型对文章整体结构与主题把握较准确，但在字数控制方面出现严重偏差。实验观察到，随着题目要求字数增加，2 个模型生成的文本超过规定字数的幅度也呈上升趋势，二者呈正相关。例如，当要求撰写 300 字左右内容时，2 个模型生成内容通常在 310 字以内；然而，当要求撰写 1 000 字的内容时，2 个模型实际生成内容可能达到1 400 字，超出要求约 40.00%，严重偏离了任务要求。这反映出当前 LLM 在生成较长文本时，仍缺乏精确有效的字数调控机制。

3.4.4　申论部分的 LLM 生成具有明显的 AI 痕迹

研究在申论测试过程中发现，LLM 生成的内容常有明显的“AI 痕迹”。尽管这类用 AI 撰写的文章通常局部段落清晰，语句通顺，但仔细审阅后会发现其具有一些共同的表达特点。例如，AI 倾向于更多使用某些公式化的连接词，如“首先”“其次”“再次”“综上所述”等。此外，即使题目的设问情景与论证的需求使作答更适合采用段落式推进写作，AI 也往往倾向于使用“第一”“第二”等方式分点阐述。

这种表达方式似乎令观点看似一目了然，但也可能牺牲论证的深度与段落间的自然过渡。AI 在生成内容时，更注重当前局部生成内容的逻辑通顺与段落的连贯，而缺少了对整篇文章的宏观布局与深层逻辑的把握。因此，AI 生成的申论内容可能看似条理分明，但细致阅读即可发现其往往缺乏人类写作中连贯的文气和论证的层层递进及整体的圆融感。

4　结语

本文以中国 NCSE 为评估场景，系统考察了主流 LLM 在此类标准化、综合性考试中的表现。研究构建了标准化提问范式，将 2022—2024 年 NCSE 真题输入各模型，通过计算正确率进行量化评估，并结合典型案例进行定性分析。

结果显示，DeepSeek–V3、DeepSeek–R1 、文心一言和 ChatGPT–4o 行测的平均正确率分别为 81.13%、83.83%、66.89%和 68.68%，均显著高于人类考生平均正确率 63.12%，表明 LLM 已具备通过此类高难度考试的潜力，在教育测评、行政辅助等领域具有应用前景。进一步比较DeepSeek 系列模型发现，DeepSeek–V3 与 DeepSeek–R1 在行测部分表现接近，未呈现显著统计学差异，说明针对推理优化的 LLM 在此次测评中未展现出压倒性优势。

研究同时揭示了当前 LLM 的局限：多数 LLM 在处理图像识别和复杂逻辑推理题目时能力不足，例如在判断推理题型中，除 DeepSeek–R1 略高于人类考生水平外，其他 LLM 均低于人类平均表现。此外，面对涉及政治敏感内容的题目时，LLM 普遍出现拒绝回答现

象,约占题库的 4%。

综上所述,LLM 在 NCSE 已展现出超越人类考生平均水平的潜力,但在图像理解、复杂推理和敏感内容处理等方面仍面临挑战。随着多模态融合、推理能力提升和领域知识整合等技术的发展,LLM 在此类复杂场景中的应用价值有望进一步增强。

参考文献:

[1] KUNG T H, CHEATHAM M, MEDENILLA A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models[J]. PLoS Digital Health, 2023, 2(2): e0000198.

[2] ZHOU M, DUAN N, LIU S, et al. Progress in neural NLP: modeling, learning, and reasoning[J]. Engineering, 2020, 6(3): 275-290.

[3] OTTER D W, MEDINA J R, KALITA J K. A survey of the usages of deep learning for natural language processing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 604-624.

[4] KATZ D M, BOMMARITO M J, GAO S, et al. GPT-4 passes the bar exam[J]. Philosophical Transactions of the Royal Society A, 2024, 382 (2270): 20230254.

[5] SKALIDIS I, CAGNINA A, LUANGPHIPHAT W, et al. ChatGPT takes on the European Exam in core cardiology: an artificial intelligence success story?[J]. European Heart Journal-Digital Health, 2023, 4(3): 279-281.

[6] TSOUTSANIS P, TSOUTSANIS A. Evaluation of large language model performance on the multi-specialty recruitment assessment (MSRA) exam[J]. Computers in Biology and Medicine, 2024, 168: 107794.

[7] XU L, CONG X, WANG R, et al. Performance of the large language models on the Chinese national nurse licensure examination: cross-sectional evaluation study[J]. JMIR medical informatics, 2025, 13: e78279.

[8] HONG M, NG W, ZHANG C J, et al. QualBench: benchmarking Chinese LLM with localized professional qualifications for vertical domain evaluation[EB/OL]. (2025-09-03)[2025-11-22]. http://arxiv.org/abs/2505.05225.

[9] 赵睿卓, 曲紫畅, 陈国英, 等. 大语言模型评估技术研究进展[J]. 数据采集与处理, 2024, 39(3): 502-523.

[10] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[EB/OL]. (2022-05-04)[2025-11-22]. http://arxiv.org/abs/2203.02155.

[11] HUANG Y, BAI Y, ZHU Z, et al. C-EVAL: a multi-level multi-discipline Chinese evaluation suite for foundation models[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2023: 62991-63010.

[12] LI H, ZHANG Y, KOTO F, et al. CMMLU: Measuring massive multitask language understanding in Chinese[C]//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024: 11260-11285.

[13] ARTIFICIAL ANALYSIS. LLM leaderboard-compare GPT-4o, Llama 3, Mistral, Gemini & other models[EB/OL]. [2025-04-16]. https://artificialanalysis.ai/leaderboards/models.

[14] HURST A, LERER A, et al. GPT-4o system card[EB/OL].(2024-10-25)[2025-04-17]. http://arxiv.org/abs/2410.21276.

[15] LIU A X, FENG B, XUE B, GOUCHER A P, et al. DeepSeek-V3 technical report[EB/OL].(2025-02-18)[2025-04-16]. http://arxiv.org/abs/2412.19437.

[16] GUO D Y, YANG D J, ZHANG H W, et al. DeepSeek-R1: incentivizing reasoning capability in LLM via reinforcement learning[EB/OL]. (2025-01-22)[2025-04-17]. http://arxiv.org/abs/2501.12948.

[17] SUN Y, WANG S H, LI Y K, et al. ERNIE: enhanced representation through knowledge integration[EB/OL]. (2019-04-19)[2025-04-17]. http://arxiv.org/abs/1904.09223.

[18] GEORGIEV P, LEI V L, BURNELL R, et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context[EB/OL]. (2024-12-16)[2025-04-17]. http://arxiv.org/abs/2403.05530.

[19] MU L L WANG X Y, JCUI J J. Evaluation of large language models for Chinese text error correction tasks[C]//The 23rd Chinese National Conference on Computational Linguistics, July 25-28, 2024, Taiyuan, China. Beijing, Chinese Information Processing Society of China, 2024: 790-806.

[20] SAHOO P, SINGH A K, SAHA S, et al. A systematic survey of prompt engineering in large language models: techniques and applications[EB/OL]. (2025-03-16)[2025-04-17]. http://arxiv.org/abs/2402.07927.

(责任编辑:刘雅婷)