

大语言模型的中国公务员考试测评

杨凯杰¹, 秦雪峰², 莫济懋², 王楚为³, 李冠霖², 丁宏强⁴, 蔡元哲²

(1. 利物浦大学, 利物浦, L69 7ZX;

2. 深圳技术大学, 大数据与互联网学院, 深圳, 518118;

3. 厦门大学马来西亚分校, 吉隆坡;

4. 香港中文大学(深圳), 深圳, 518172)

摘要: 研究对包括 Deepseek-R1 在内的多种大型语言模型在中国国家公务员考试(简称“国考”)中的应试能力进行了系统评估与分析。研究选取 2022—2024 年中国国家公务员考试真题, 以预设的标准化提问范式向 ChatGPT-4o、Deepseek-V3 和 Deepseek-R1 三种模型分别输入题目并记录其输出结果, 进而统计各模型的答题正确率以衡量综合能力。实验结果显示, ChatGPT-4o 的综合正确率为 68.68%, Deepseek-V3 为 81.13%, Deepseek-R1 为 83.83%, 均高于人类考生平均正确率 63.12%。为深入了解各模型的优势与不足, 还对判断推理、数量关系、言语理解与表达、资料分析等不同题型的答题表现进行了细分对比, 并归纳了各模型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; 公务员考试; 性能评估

Chinese Civil Service Examination Assessment Using Large Language Models

YANG Kaijie¹, QIN Xuefeng², MO Jimao², WANG Chuwei³, LI Guanlin², DING H.Q. Chris⁴, CAI Yuanzhe²

(1. Universit of Liverpool, Liverpool, L69 7ZX, UK;

2. College of Big Data and Internet, Shenzhen University of Technology, Shenzhen, 518118, China;

3. Xiamen University Malaysia, Kuala Lumpur;

4. The Chinese University of Hong Kong (Shenzhen), Shenzhen, 518172, China)

Abstract: The study systematically evaluated and analyzed the test-taking ability of various large-scale language models, including Deepseek R1, in the Chinese National Civil Service Examination (abbreviated as "National Examination"). The study selected the real questions of the Chinese National Civil Service Examination from 2022 to 2024, and input the questions to the three models of ChatGPT-4o, Deepseek-V3 and Deepseek-R1 respectively with a preset standardized questioning paradigm and recorded their output results, and then counted the correct answer rate of each model to measure the comprehensive ability. The experimental results show that the comprehensive correct rate of ChatGPT-4o is 68.68%, Deepseek-V3 is 81.13%, and Deepseek-R1 is 83.83%, all of which are higher than the average correct rate of 63.12% of human candidates. In order to gain a deeper understanding of the advantages and disadvantages of each model, the performance of answering different question types such as judgment reasoning, quantitative relationship, verbal comprehension and expression, and data analysis was also subdivided and compared, and the typical error types of each model when dealing with complex logical reasoning and multi-step operation questions were summarized.

Keywords: large language model; civil service examination; performance evaluation

大型语言模型 (Large Language Models, LLMs) 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领域, 更被视为迈向通用人工智能 (Artificial general

intelligence, AGI) 的关键一步^[1-4]。这些强大的模型已应用至教育、医疗、法律等多个行业, 展现出解决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^[5]。与此同时, LLMs 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 LLMs 在如美国医学执业资格考试 (United States Medical Licensing Examination, USMLE) 等专业测试中取得了令人瞩目的成绩^[6]。

在此背景下, 中国国家公务员考试 (National Civil Servant Exam, NCSE) 为评估 LLMs 的能力提供了一个独特且重要的评估场景。近年来, 随着高等教育普及化和高校毕业生数量激增 (2022 年已达 1076 万), 就业竞争日趋激烈, 使得公务员成为备受追捧的职业发展路径, 其社会关注度和重要性日益提升^[7,8]。中国公务员考试主要由《行政职业能力

测验》(以下简称“行测”) 与《申论》两部分组成, 前者强调逻辑推理、数据分析与语言理解等综合能力, 后者则聚焦于政策分析、写作与综合表达能力。这些对考生综合素质的极高要求, 也对 LLMs 的能力构成了严峻的考验。

LLMs 在适应中国特有文化背景、复杂语境及高难度要求的公务员考试方面的能力尚不明确。因此, 本研究旨在系统评估以 Deepseek-R1 为代表的主流大型语言模型在 2022 年至 2024 年中国国家公务员考试中的表现。我们不仅关注模型是否能达到“通过”标准, 更致力于深入分析不同模型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。本研究的核心目标是回答: 当前的大型语言模型是否具备通过中国公务员考试的能力? 其表现揭示了 LLMs 在应对复杂、本土化标准化测试时的优势与局限。

1 相关工作

大语言模型是能够理解和生成人类语言的计算模型^[5]。大语言模型 (LLMs) 的飞速发展, 已显著提升了自然语言处理 (Natural Language Processing, 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^[9,10]。这些模型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 LLMs 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 LLMs 应用于各类标准化考试的评估中, 考察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 LLMs 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, Bommarito 等人证实 ChatGPT-4o 能够通过美国律师资格考试^[11]; Skolidis 等人表明 ChatGPT-4o 在欧洲核心心脏病学考试中获得了及格成绩^[12]; Tsoutsanis 等人的研究显示, 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^[13]。上述研究充分展示了大型语言模型在处理高难度专业语言任务中的能力, 激发了我们对其在中国公务员考试等高复杂度、文化语境依赖性强的标准化测试中表现的关注。

中国国家公务员考试以其独特的复杂性、对精微语言理解和文化背景知识的要求而著称, 加之不同 LLMs 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, 使得它们在 NCSE 上的适用性成为一个亟待验证的问题^[14]。尽管 LLMs 在部分标准化测试中已展现潜力^[11-13], 但针对 NCSE 这一复杂场景, 特别是比较不同模型表现的系统性研究尚属空白。因此, 本研究旨在核心评估当前代表性 LLMs (如 Deepseek-R1) 能否应对 NCSE 的综合要求, 并关键性地对比分析这些具有不同特性的模型之间的能力差异, 从而为理解 LLMs 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 方法

2.1 数据集

本研究选取了 2022 至 2024 年间国家公务员考试的完整试题作为评估材料, 该试题来源于公开的公务员考试辅导平台粉笔网^[15]。其中, “行测”部分共包含 130 道单项选择题, 按能力模块划分为五个子部分: 常识判断、言语理解与表达、数量关系、判断推理以及资料分析。

“行测”部分题型覆盖内容及考察方向具体如下: 常识判断: 涵盖人文、科技、法律、历史等多学科领域, 考察考生的基础知识储备与综合素养; 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; 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; 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; 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”各部分的题型具体结构与题量分布如表 1 所示。

“申论”部分则采用材料写作形式, 由五则材

料及对应的五道主观题组成。主要考察考生对给定材料的信息提取、概括、梳理、归纳与提炼能力，

并要求考生能够基于材料提出观点、进行论证，最终形成结构完整、论证充分的议论文。

表1 中国国家公务员考试试卷组成

Table 1 Composition of the Chinese National Civil Servant Examination Paper

行测题型构成	分数构成	题目数量	答题时间
常识判断	10	20	2 小时
言语理解与表达	30	40	
数量关系	15	15	
判断推理	25	35	
资料分析	20	20	3 小时
申论	100	5	

2.2 参评模型

本研究旨在对当前具有较高影响力及先进性的大语言模型进行中国公务员考试评估。模型的选取主要遵循以下原则：(1) 模型在学术界与产业界具有广泛的关注度和应用基础；(2) 模型在公开的评测基准或相关研究中展现出领先或具有代表性的性能^[16]；(3) 模型的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^[17-21]。除上述因素外，本研究特别考虑了国内大模型，因为国内大模型通常会使用大量具有中国特色，中文的数据集进行训练，而海外大模型则可能会选取全球范围内的数据资源^[22]。

基于上述原则，本研究最终确定纳入评估的模型共计五款。具体包括：ChatGPT-4o^[17]，Gemini-1.5 Flash^[21]（以下简称 Gemini），ERNIE Bot-4.0 Turbo^[20]（以下简称文心一言），Deepseek-V3^[18]与 Deepseek-R1^[19]。各模型基本信息如表2所示。

表2 参评模型信息

Table 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 ^[17]
Gemini-1.5 Flash	Google	2024.09	记忆力强（支持超长对话），响应快，性价比高 ^[21]
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 ^[20]
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 ^[18]
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 ^[19]

2.3 提问范式

既有研究已充分证明，提示词（Prompt）的设计对大型语言模型的输出质量和行为具有显著影响^[23]。

为确保评测结果的稳定性，减少因提示词差异引入的实验偏差，本研究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参见图1所示。

行测部分示例提问：
请回答下面的单选题

——提示词

81.单选题

同种数罪是指行为人实施数个行为，符合数个性质相同的基本犯罪构成，触犯数个罪名相同的数罪。

根据上述定义，下列情形属于同种数罪的是：

- A.甲为了自杀，在水中投入剧毒农药，其幼女误饮后身亡，甲见状也饮水身亡
- B.乙为阻止警察抓捕其子，采用暴力手段妨害警察的执法行为，挥刀将警察砍成重伤
- C.丙因缺钱，先是对某单位实施了盗窃，而后又使用暴力手段抢劫了素不相识的赵某
- D.丁故意杀死了与其分手的女友张某，而后又杀死了在生意上与其竞争的李某

——题目原文

申论部分示例提问：

“给定资料1”反映了N市积极落实惠企政策的有关情况，请简述其主要做法及成效。（15分）

要求：全面、准确、有条理。不超过250字。

新型冠状病毒肺炎疫情发生后，党中央、国务院出台了一系列支持企业发展的惠企政策。N市积极落实各项惠企政策，不断优化营商环境，推动区域经济高质量跨越式发展。

“当时举步维艰，那笔钱可真是雪中送炭啊。”宏远环保科技有限公司石经理现在回想起来，仍然有些感慨。疫情影响下，这家公司资金严重短缺，已经上马的技改项目濒临停滞。2020年7月，该公司通过N市“政策直达直兑”绿色通道申请了两项技改专项补贴和一项专利奖励……（申论考试材料原文全文）

图1 提问范式

Figure 1 Questioning Paradigm

2.4 测试步骤

本研究采用标准化输入流程，将规范化试题以单

题为单位逐次输入网页端大语言模型进行测试。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，待模型完成响应后立即保存输出记录；随后通过启动独

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。该方法设计确保各测试单元间具有独立性，降低情境线索对模型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

在获得大语言模型对全部试题的作答结果后，研究团队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，本研究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五大模块，分别统计各模型在不同模块的得分分布特征。该分析框架不仅能够量化模型的能力异质性特征，还可直观反映其在标准化考试场景中的优势与局限。同时，本研究统计了各模型在各类题型中的准确率（Accuracy Rate）指标，其数学表达式如下：

$$r = \frac{\alpha - \beta}{\alpha} 100$$

其中， r ：模型正确率； α ：答题总数； β ：错误回答题目数；

2.6 申论专家评分标准

为确保对模型生成的申论内容进行专业且客观的评估，本研究特邀深圳技术大学马克思主义学院的专家学者承担评分工作。这些专家在马克思主义理论、公共策论及申论研究领域具备深厚学术背景与丰富的校内评审经验。其评估将严格参照既有的申论评价规范，并结合校内教学产生的评估标准，对模型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 《行政能力测试》结果分析

3.1.1 常识判断

常识判断中 ChatGPT-4o、Gemini、文心一言、Deepseek-V3 和 Deepseek-R1 的总体正确率分别是 89.47%、77.78%、84.21%、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，因此在统计上述最终正确率时我们剔除了这部分题目。为了多维度评估大语言模型在常识判断任务中的能力，本研究使用“粉笔”网站试题分类，对模型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图表 3 所示。

为具体在所测试的三年题目中，ChatGPT-4o 共获得 89.47%的正确率，整体表现优异，尤其在政治常识类题目中取得了 100%的较高正确率，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 模型的总正确率为 77.78%，表现较为一般，刚刚到达及格表现。文心一言 获得较好正确率，为 84.21%。涵盖多个知识领域，包括技术、人文与法律等，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。

综合各模型的得分情况可见，不同模型在通识知识处理能力方面存在显著差异。Deepseek 的两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。ChatGPT-4o 紧随其后，在多领域表现出色，具有较强的泛化能力。Gemini 的两个模型则表现稍逊于 ChatGPT-4o 和 Deepseek 的两个模型。相较之下，文心一言的表现则相对有限，其在某些专业领域的知识掌握深度与答题策略仍有待优化。

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，但同时，Deepseek-R1 更容易出现拒绝回答题目的现象，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部内容安全审查机制，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

并非完全反映模型的基础推理能力，甚至提示 Deepseek-V3 在此特定场景下具备相对优势。

常识判断部分的测试结果揭示：在记忆性知识的精确掌握方面，人类表现与大语言模型之间存在显著差距。鉴于大语言模型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，可以考虑引入大语言模型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。

表 3 常识判断题型模型正确率（2022-2024）

Table 3 Model Accuracy Rate on Common Sense Judgment Question Types (2022-2024)

题型分类（对应题数）	ChatGPT-4o	Gemini	文心一言	Deepseek-V3	Deepseek-R1	人类考生
政治常识	100%	66.67%	100%	88.89%	100%	52.68%
经济常识	100%	100%	100%	100%	100%	34%
科技常识	33.33%	33.33%	66.66%	100%	100%	41.53%
人文常识	50%	0%	100%	100%	100%	47.45%
地理常识	100%	100%	100%	100%	100%	38.40%
法律常识	85.71%	90%	85.71%	95%	90%	38.89%
总体正确率	89.47%	77.78%	84.21%	95.92%	97.78%	40.30%

3.1.2 言语理解与表达

模型测试结果如表 4 所示，表中列出了 2022-2024 年五种主流大语言模型在“言语理解与表达”模块下各子题型中的平均作答情况。表格最后一行展示了各模型在该模块下的总正确率，并附有人类考生的平均正确率作为参照标准。结果显示，参评的五种模型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。

其中，Deepseek-R1 与 Deepseek-V3 的总正确率分别为 88.3%与 87.5%，在所有模型中排名靠前，表明它们在处理中文语言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。其余三种模型的正确率从高到低依次为：ChatGPT-4o（73.3%）、文心一言（73.3%）与 Gemini（60%）。除 Gemini 以外，其余模型的总体表现均超越了人类考生的平均水平（69.2%），这凸显了它们在语义分析与语言逻辑推理方面已具备的坚实能力。此外，表 4 还详细列出了各模型在 2022-

2024 年不同年度的言语理解与表达及其具体子题型上的得分，为深入分析其在不同时间及任务类型下的性能稳定性与表现差异提供了数据支持。

总体而言，大语言模型在处理“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，Deepseek 的两个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。对于大模型之间这样的不同表现，本研究得出了以下几点：1. 训练语料的规模 2. 训练方向的差异 3. 训练强度与策略差异 综上，模型在“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。

表 4 言语理解与表达题型模型平均正确率（2022-2024）

Table 4 Average Model Accuracy Rate on 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	人类考生平均
逻辑填空	68.35%	51.65%	88.35%	86.65%	58.35%	数据暂缺
片段阅读	69.73%	69.73%	90.91%	87.91%	87.91%	数据暂缺
语句表达	88.89%	74.11%	85.22%	88.89%	88.89%	数据暂缺

总体正确率	73.3%	60%	88.3%	87.5%	73.3%	69.2%
-------	-------	-----	-------	-------	-------	-------

3.1.3 数量关系

表5展示了五种主流大语言模型在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，AI模型亦呈现出显著的性能波动。表中列出了Deepseek-R1、文心一言、ChatGPT-4o、Deepseek-V3与Gemini在2022至2024年“行测”考试中数量关系部分的正确率变化情况。

从结果来看，Deepseek-R1整体表现最为优异，特别是在2022年的试题中，均取得各模型中的最高

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，多数模型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，虽然大型语言模型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。进一步提升模型的数学建模能力、上下文解析能力以及推理链条的完整性，是未来研究与优化的关键方向。

表5 数量关系类别目各小题模型正确率

Table 5 Model Accuracy Rate for Each Individual Question, by Quantitative Relations Category and Sub-item

题型年份 (对应题数)	ChatGPT-4o	Gemini	文心一言	Deepseek-V3	Deepseek-R1	人类考生
22年正确率	50.00%	20.00%	40.00%	80.00%	80.00%	数据暂缺
23年正确率	60.00%	40.00%	30.00%	60.00%	70.00%	数据暂缺
24年正确率	70.00%	30.00%	50.00%	50.00%	70.00%	数据暂缺
平均正确率	60.00%	30.00%	40.00%	63.33%	73.33%	38.73%

3.1.4 判断推理

在公务员考试判断推理模块的各个子任务中，分别有如图2的正确率。不同大语言模型（LLMs）展现出显著的能力差异。

DeepSeek系列模型（R1与V3版本）在此次评估中表现最优，其中DeepSeek-R1以70.85%的总体正确率位居榜首，DeepSeek-V3则以69.038%的正确率紧随其后。值得注意的是，在定义判断任务30小题中，有3道小题存在图形信息，导致Deepseek模型无法正常作答，我们在处理结果时剔除了这部分题目。在此基础上，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%、86.7%和96.7%，；DeepSeek-V3在三项任务中的正确率依次为96.15%、86.7%和93.3%。然而，由于当前版本缺乏图像处理能力，两款模型在30道图形推理题目中均未得分。尽管存在这一能力缺口，其文本推理优势仍使其综合表现优于其他模型。

文心一言以62.5%的总体正确率排名第三，其在图形推理和类比推理子任务中的正确率分别为53.3%和56.7%。该模型在定义判断与逻辑判断任务中分别

取得73.3%和67.3%的正确率。相较于其他模型，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57.5%的总正确率位列第四，在定义判断任务中表现出色（80.0%正确率），但在图形推理任务中仅获30.0%正确率，构成其整体表现的主要短板。谷歌Gemini模型表现相对较弱，四项任务总平均正确率为43.3%，其中图形推理任务得分最低（20.0%），定义判断任务得分最高（56.7%）。

值得注意的是，人类考生在该模块的平均总正确率为69.35%（其中图形推理72.5%、逻辑判断66.3%）。DeepSeek-R1（70.85%）的综合得分超出人类基准线，其余模型均低于人类水平。人类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%阈值，且个体间差异幅度较小。

研究显示，虽然部分先进LLMs（尤其是DeepSeek系列）在公务员考试判断推理模块已具备超越人类的文本推理能力，但其能力图谱呈现明显的非对称性特征——所有模型在图形推理维度均存在系统性缺陷。这种认知能力的不完整性揭示出当前LLMs在模拟人类多模态推理机制方面仍面临重大技术挑战。

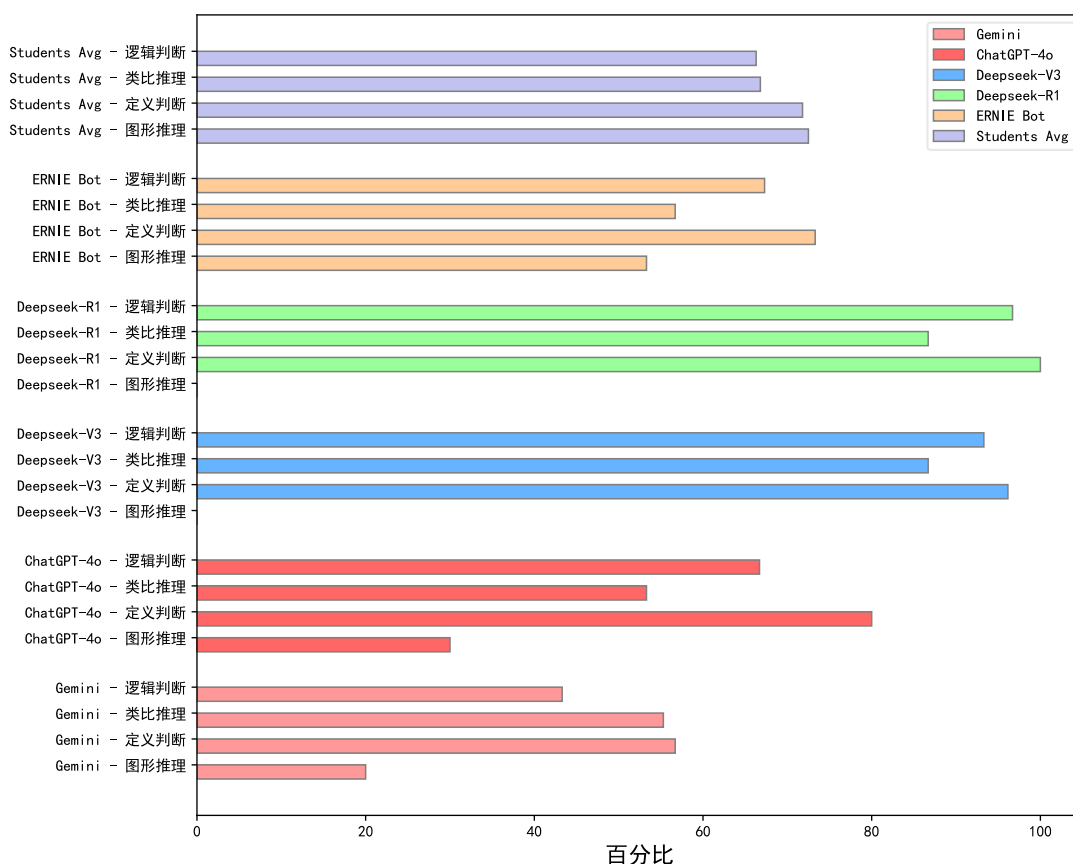


图 2 判断推理部分各模型正确率数据

Figure 2: Accuracy Rate Data for Each Model on the Judgment and Reasoning Section

3.1.5 资料分析

在“资料分析”模块的测试中，各大语言模型在处理涉及图表理解、数据信息提取与综合计算类题型时的正确率如表 6 所示。具体而言，Deepseek-R1 与 Deepseek-V3 分别在可处理的 15 道纯文本题目子集上取得了 100%和 93.3%的极高正确率；然而，

由于在本测试环境下缺乏处理图像信息的能力，它们未能完成包含图表在内的其余 45 道题目。与之对比，具备多模态处理能力的 ChatGPT-4o、文心一言及 Gemini 模型则尝试了全部 60 道题目，其涵盖图表理解、数据提取与计算的整体平均正确率分别为 61.5%、30.0%和 26.5%。人类考生在该完整模块上的平均正确率为 72.9%。

表 6 2022-2024 年资料分析部分各模型得分情况

Table 6 Scoring Performance of Each Model in the Data Analysis Section 2022-2024

模型	准确率(2022)	准确率(2023)	准确率(2024)	平均准确率
ChatGPT-4o	55%	65%	65%	61.5%
文心一言	35%	20%	35%	30%
Gemini	30%	30%	20%	26.5%
Deepseek-V3	80%	100%	100%	93.3%
Deepseek-R1	100%	100%	100%	100%
人类考生	73.05%	73%	72.7%	72.9%

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。而其他模型（尤其是 ChatGPT-4o）虽然

整体得分较低，但其分数代表了在更全面的、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模型间在数据分析能力上的差距，

更凸显了多模态能力对于完成真实世界复杂数据分析任务的重要性,并表明当前直接比较不同能力范围模型得分的做法需格外审慎。

3.1.6 实验小结

最后,根据上述数据,可得出各模型在 2022–2024 三年行测测试题中的平均正确率,如表 7 所示。

表 7 各模型 2022–2024 年行测平均正确率

Table 7: Average Accuracy Rate of Each Model on the Administrative Aptitude Test (2022–2024)

模 型	Gemi ni	Deepse ek-V3	Deepse ek-R1	文心 一言	ChatGP T-4o	人类 考生
平 均 正 确 率	48.80 %	81.13%	83.83%	66.89 %	68.68%	63.12 %

3.2 申论

在“申论”模块的测试中,各模型在申论考试中不同材料的的具体得分如表 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

表 8 中国公务员考试申论部分的各模型小题平均得分(2022–2024)

Table 8 Average scores of each model question in the essay section of the Chinese civil service examination (2022–2024)

题目	ChatGPT-4o	Deepseek-R1	Gemini	Deepseek-V3	文心一言	人类考生(平均/最高)
材料 1	5.33	5.83	4.00	6.77	0.33	数据暂缺
材料 2	7.17	9.00	6.17	10.33	2.17	数据暂缺
材料 3	10.17	10.50	10.17	11.50	4.50	数据暂缺
材料 4	14.33	12.50	13.33	11.67	5.33	数据暂缺
材料 5	24.17	22.33	24.17	23.00	13.17	数据暂缺
平均得分	61.17	60.16	57.84	63.27	25.50	30.38/87

3.3 结果比较与案例分析

在对多个大语言模型进行系统测试后,研究团队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,部分模型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

其中 Deepseek 系列模型取得了最高得分,Deepseek-R1 得分最高。在表中可以观察到大部分的大语言模型的答题成绩要由于人类考生,说明大语言模型在行测考试方面是优于人类考生的。

任务中取得了 63.27 的平均分,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 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 的表现材料理解、信息整合与文字表达方面已展现出较强的能力。

尽管 Deepseek-V3 取得了令人鼓舞的成绩,但其表现在与顶尖人类考生的比较中,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、语言运用的灵活性以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,不仅反映了当前大语言模型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,大语言模型已能提供有力的支持。

知偏差与系统性误差。

3.3.1 大型语言模型作答结果和推理结果有可能不同:

在例题 1 中,题目需要我们计算一个沿着圆内接正方形边垂直切掉的部分的表面积。大语言模型选择的是 D 选项,但题目实际答案是 C。可以看出在最后一步中,GPT 的结果和推理是两个不同的过程。 $(S_{\text{切去部分}} = (\pi + (2^{1/2}) * \pi) - 24)$ 是 ChatGPT-4o 计算出的结果,但其最后的作答结果是(S 切去部分

$=2((2^{1/2})\pi - 2)$ 。可以看到这两个结果不匹配，因此说明大模型作答结果与推理结果是不同的。

例题 1：圆柱体的高度为 1，底面内接正方形边长为 1。

圆柱体被切割 4 次，切割后得到棱长为 1 的正方体，需计算切割去部分的总表面积。

- A. $(2^{1/2})\pi + 2$
 B. $2(2^{1/2})(\pi - 2)$
 C. $((2^{1/2}) + 1)\pi + 2$
 D. $2(2^{1/2})\pi - 2$

GPT 解题步骤：

根据题目，圆柱体底面内接正方形的边长为 1，则圆柱体底面半径为：

$$r = (2^{1/2}) / 2$$

高度 $h = 1$ 。

计算圆柱体的总表面积：

$$S_{\text{圆柱}} = 2 * \pi * ((2^{1/2}) / 2)^2 + 2 * \pi * ((2^{1/2}) / 2) * 1$$

$$S_{\text{圆柱}} = \pi + (2^{1/2}) * \pi$$

每个正方体的棱长为 1，总表面积为： $S_{\text{正方体}} = 6 * 1^2 = 6$

切割后得到 4 个正方体，则总表面积为：

$$S_{\text{总正方体}} = 4 * 6 = 24$$

切去部分的表面积为原圆柱的表面积减去 4 个正方体的总表面积：

$$S_{\text{切去部分}} = S_{\text{圆柱}} - S_{\text{总正方体}}$$

$$S_{\text{切去部分}} = (\pi + (2^{1/2}) * \pi) - 24$$

根据选项形式进行匹配，最终结果为：

$$S_{\text{切去部分}} = 2(2^{1/2})\pi - 2$$

因此选择：D

3.3.2 大模型推理会由于图片识别产生偏差：

此外，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大语言模型的核心能力边界。在例题 2 中，题目需要考生找出图形之间的规律，并正确分类。参评模型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大语言模型在处理多模态信息，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，在如图形算术或资料分析等题型中，模型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。

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，模型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，另一种常见情况是，模型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，大语言模型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，构成了当前大语言模型在应对包含视觉元素的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，也明确指出了未来模型迭代升级中，在提升多模态理解与推理能力方面亟待加强的方向。

例题 2：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：

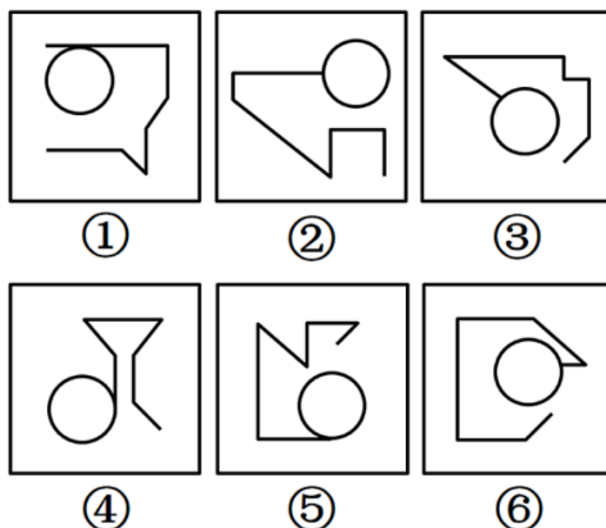


图 3 大语言模型图片识别案例

Figure 3 Large language model image recognition case

- A. ①②④, ③⑤⑥
 B. ①③⑤, ②④⑥
 C. ①④⑤, ②③⑥
 D. ①②⑥, ③④⑤

正确答案：C。

Deepseek-V3 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。

3.3.3 政治敏感的问题大模型会拒绝回答：

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，在例题 3 中，题目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。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 模型在行测的部分题型中也出现了这种现象。其原因可能是模型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，这些机制导致其在特定情况下无法做出实质性回答，下面的例题可以反映这个现象。值得一提的是，经过测试，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。

尽管这些内容在实际考试环境中通常并不构成敏感话题，部分模型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，也为未来大语言模型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，提升模型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性，将成为后续模型设计中亟需权衡的关键问题。

例题 3：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。

3.3.4 大型语言模型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，部分大语言模型在处理某些任务时，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 的两个模型进行深入分析，结果显示，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，模型生成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，模型生成内容通常在 310 字以内；然而，当要求撰写 1000 字的内容时，模型实际生成内容可能达到 1400 字，超出要求约 40%，严重偏离了任务要求。这清晰反应出当前大语言模型在生成较长文本时，仍缺乏精确有效的字数调控机制。

3.3.5 申论部分的大语言模型生成具有明显的 AI 痕迹：

研究在申论测试过程中发现，大语言模型生成的内容常有明显的“AI 痕迹”。尽管这类有 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它们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、论证的层层递进以及整体的圆融感。

4 结论

本研究以中国国家公务员考试为评估场景，系统考察了主流大语言模型（LLMs）在应对此类标准化、综合性考试时的表现。研究首先构建了一套标准化的提问范式，并依据此范式将 2022 至 2024 年间的国家公务员考试真题逐一输入至参评的各款 LLMs。随后，通过记录各模型的输出结果并计算其在各考试模块的答题正确率，对它们的整体应试表现进行了量化评估。同时，本研究还结合部分典型作答案例，对模型生成答案的内容质量及其内在的作答逻辑展开了深入的定性分析。

研究结果显示，Deepseek-V3，Deepseek-R1 与 ChatGPT-4o 分别取得了 81.13%，83.83%和 68.68%的总正确率，均显著高于所参照的人类考生平均正确率（63.12%）。这一发现表明，现有的大语言模型在应对如国家公务员考试这类高难度标准化测试时已展现出一定的应用潜力，并初步证明其具备了通过此类考试的能力。这些结果预示着大语言模型在未来的教育测评、行政辅助以及初步人才筛选等场景中拥有广阔的发展前景。

本研究进一步对深度求索（Deepseek）系列的两款不同模型（Deepseek-V3 与 Deepseek-R1）的性能进行了比较。实验结果显示，在“行测”部分的整体测试中，Deepseek-V3 的平均正确率为 81.13%，而 Deepseek-R1 的平均正确率为 83.83%。考虑到实验设计中可能存在的偶然误差及两者百分比差异的幅度，可以认为在本研究的特定测试条件下，这两款模型在“行测”部分的整体表现并未呈现出统计学意义上的显著差距。因此，一个初步的结论是，至少在此次公务员“行测”部分的评估中，针对推理能力进行特定优化的模型相较于其对应的基础模型，并未展现出压倒性的性能优势。

同时，本研究也揭示了当前 LLMs 在应对此类考试时存在的若干局限。多数参评模型在处理涉及图像识别的题目时能力不足，并且在某些复杂逻辑推理任务中表现欠佳。以“判断推理”模块为例，文心一言（62.5%）、ChatGPT-4o（57.5%）和

Gemini (43.3%) 的得分均低于人类考生的平均水平 (69.35%), Deepseek-V3 的正确率为 69.038%, 而仅 Deepseek-R1 (70.85%) 略高于人类平均表现。这在一定程度上反映出这些模型在形式逻辑的严密性和语义一致性的深层建模方面尚有提升空间。此外, 在面对包含潜在政治敏感内容的题目时, 不同 LLMs 均出现了拒绝回答的现象, 此类题目平均约占本次测试题库总量的 4%。针对 Deepseek 系列模型的特定情况, 其当前版本主要具备图片内文字识别能力, 而对于测评中出现的抽象图形推理题则无法进行有效作答, 这类因模型能力所限而未能作答的题目平均占比达到了总题量的 5%。

尽管本研究提供了一些有价值的初步洞见, 但其本身仍存在若干局限。其一, 当前的测评数据仅基于近三年中国国家公务员考试的公开题库, 虽然具有一定的代表性, 但样本数量相对有限, 可能难以完全且无偏地反映出各模型在所有潜在考试情境下的真实能力边界。其二, 本次纳入测评的 LLMs 数量亦较为有限, 虽已包含若干主流及代表性模型, 但尚未能覆盖当前市场上所有重要的大语言模型的完整谱系。

为进一步深化对 LLMs 在公务员考试等复杂应用场景中能力的理解与评估, 未来研究可在以下几个关键方向进行扩展与深化: 首先, 应致力于扩大数据集的规模与多样性, 例如引入更多考试年度、覆盖更广泛地区的公务员考试题目, 并特别关注对包含图表、图像等跨模态信息题目的处理能力的测试。其次, 需要拓展并丰富多维度的评估指标, 在目前主要依赖的正确率指标之外, 应探索并结合诸如模型作答的平均响应时间、答案输出的稳定性与一致性、内部逻辑推理链条的完整性与准确性, 以及答案的可解释性等多个维度, 以构建一个更为全面、细致的性能评价框架。最后, 应加强对模型特定语境适应能力的深度测试, 特别是针对公务员考试中常见的、富含中国特有文化背景、复杂政策法规以及特定法律条文解读的题目, 深入考察模型的理解适应能力及其内置内容安全策略在处理此类题目时的具体表现和潜在影响。

综上所述, 当前的大语言模型在应对中国国家公务员考试这类综合性、高难度标准化测试时, 已经展现出不容忽视的应用潜力, 部分顶尖模型的表现甚至超越了人类平均水平。尽管如此, 它们在图像理解的深度、复杂逻辑推理的严谨性以及特定敏感话题的恰当处理等方面仍面临显著挑战。我们有理由预期, 随着人工智能技术的持续快速发展与迭代, 特别是在多模态信息融合技术、高级推理能力的构建以及对特定领域知识的深度整合与理解等方面的不断优化与突破, 大语言模型在此类复杂应用场景中的作用与价值必将得到进一步的增强与拓展。

5 基金资助

该项目获得深圳市高校稳定项目 (20231127194506001) 和广东省高校创新项目 (2024KTSCX055) 支持。

参考文献:

- [1] KASAI J, KASAI Y, SAKAGUCHI K, 等. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations[A/OL]. arXiv, 2023[2024-12-03]. <http://arxiv.org/abs/2303.18027>. DOI:10.48550/arXiv.2303.18027.
- [2] SAVALÉ N. Quantitative version of Weyl's law[A/OL]. arXiv, 2023[2025-04-11]. <http://arxiv.org/abs/2303.01407>. DOI:10.48550/arXiv.2303.01407.
- [3] EDDINE N C, REPOVŠ D D. The Neumann problem for a class of generalized Kirchhoff-type potential systems[J/OL]. Boundary Value Problems, 2023, 2023(1): 19. DOI:10.1186/s13661-023-01705-6.
- [4] BUBECK S, CHANDRASEKARAN V, ELDAN R, 等. Sparks of Artificial General Intelligence: Early experiments with GPT-4[A/OL]. arXiv, 2023[2025-04-15]. <http://arxiv.org/abs/2303.12712>. DOI:10.48550/arXiv.2303.12712.
- [5] CHANG Y, WANG X, WANG J, 等. A Survey on Evaluation of Large Language Models[A/OL]. arXiv, 2023[2025-04-11]. <http://arxiv.org/abs/2307.03109>. DOI:10.48550/arXiv.2307.03109.
- [6] KUNG T H, CHEATHAM M, MEDENILLA A, 等. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models[J/OL]. PLOS Digital Health, 2023, 2(2): e0000198. DOI:10.1371/journal.pdig.0000198.
- [7] 张仲琪, 高翠欣. 后疫情时代大学生“考公务员热”现象研究[J/OL]. 中国大学生就业, 2022(14): 11-17. DOI:10.20017/j.cnki.1009-0576.2022.14.002.(Zhang Zhongqi, Gao Cuixin. Research on the Phenomenon of College Students' Fever for Civil Servants in the Post-epidemic Era[J/OL]. Journal of Chinese College Student Employment, 2022(14): 11-17)
- [8] 吴奕. 为公务员热降温还需哪几招[J]. 人民论坛, 2018(11): 54-55.(Wu Yi. What more measures are needed to cool down the civil servant fever?[J]. People's Tribune, 2018(11): 54-55.)
- [9] ZHOU M, DUAN N, LIU S, 等. Progress in Neural NLP: Modeling, Learning, and Reasoning[J/OL]. Engineering, 2020, 6(3): 275-290. DOI:10.1016/j.eng.2019.12.014.
- [10] OTTER D W, MEDINA J R, KALITA J K. A Survey of the

Usages of Deep Learning for Natural Language Processing[J/OL]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 604-624. DOI:10.1109/TNNLS.2020.2979670.

[11] KATZ D M, BOMMARITO M J, GAO S, 等. GPT-4 passes the bar exam[J/OL]. Philosophical Transactions of the Royal Society A, 2024[2025-04-11]. <https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0254>. DOI:10.1098/rsta.2023.0254.

[12] SKALIDIS I, CAGNINA A, LUANGPHIPHAT W, 等. ChatGPT takes on the European Exam in Core Cardiology: an artificial intelligence success story?[J/OL]. European Heart Journal - Digital Health, 2023, 4(3): 279-281. DOI:10.1093/ehjdh/ztad029.

[13] TSOUTSANIS P, TSOUTSANIS A. Evaluation of Large language model performance on the Multi-Specialty Recruitment Assessment (MSRA) exam[J/OL]. Computers in Biology and Medicine, 2024, 168: 107794. DOI:10.1016/j.combiomed.2023.107794.

[14] 赵睿卓, 曲紫畅, 陈国英, 等. 大语言模型评估技术研究进展[J/OL]. 数据采集与处理, 2024, 39(3): 502-523. DOI:10.16337/j.1004-9037.2024.03.002.(Zhao Ruizhuo, Qu Zichang, Chen Guoying et al. Research Progress in Evaluation Techniques for Large Language Models[J/OL]. Journal of Data Acquisition and Processing, 2024, 39(3): 502-523)

[15] 公务员考试、事业单位、教师、司法、考研、建造和会计等考试培训[EB/OL]. [2025-04-13]. <https://www.fenbi.com/>(Examination training for civil service examinations, public institutions, teachers, judicial examinations, postgraduate entrance examinations, construction and accounting examinations[EB/OL],2025-04-13, <https://www.fenbi.com/>)

[16] LLM Leaderboard - Compare GPT-4o, Llama 3, Mistral, Gemini & other models | Artificial Analysis[EB/OL]. [2025-04-16]. <https://artificialanalysis.ai/leaderboards/models>.

[17] OPENAI, HURST A, LERER A, 等. GPT-4o System Card[A/OL]. arXiv, 2024[2025-04-16]. <http://arxiv.org/abs/2410.21276>. DOI:10.48550/arXiv.2410.21276.

[18] DEEPSEEK-AI, LIU A, FENG B, 等. DeepSeek-V3 Technical Report[A/OL]. arXiv, 2025[2025-04-16]. <http://arxiv.org/abs/2412.19437>. DOI:10.48550/arXiv.2412.19437.

[19] DEEPSEEK-AI, GUO D, YANG D, 等. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning[A/OL]. arXiv, 2025[2025-04-16]. <http://arxiv.org/abs/2501.12948>. DOI:10.48550/arXiv.2501.12948.

[20] SUN Y, WANG S, LI Y, 等. ERNIE: Enhanced Representation through Knowledge Integration[A/OL]. arXiv,

2019[2025-04-17]. <http://arxiv.org/abs/1904.09223>. DOI:10.48550/arXiv.1904.09223.

[21] TEAM G, GEORGIEV P, LEI V I, 等. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context[A/OL]. arXiv, 2024[2025-04-17].

<http://arxiv.org/abs/2403.05530>. DOI:10.48550/arXiv.2403.05530.

[22] LINGLING M, XIAOYING W, JIAJIA C. 大语言模型在中文文本纠错任务的评测(Evaluation of large language models for Chinese text error correction tasks)[C/OL]//SUN M, LIANG J, HAN X, 等. Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference). Taiyuan, China: Chinese Information Processing Society of China, 2024: 790-806[2025-04-17]. <https://aclanthology.org/2024.ccl-1.62/>.

[23] SAHOO P, SINGH A K, SAHA S, 等. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications[A/OL]. arXiv, 2025[2025-04-17]. <http://arxiv.org/abs/2402.07927>. DOI:10.48550/arXiv.2402.07927.

