

大语言模型的中国公务员考试测评

杨凯杰, 李冠霖, 王楚为, 莫济懋, 蔡元哲, 黄飞娟, 丁宏强

摘要

【目的】 本研究旨在评估 ChatGPT-4o 等大语言模型在中国公务员考试中的表现, 探讨其在复杂、高风险标准化考试中的应用前景与局限性。**【方法】** 选取三年公务员考试试题, 通过规定标准提问范式, 输入大语言模型并记录输出结果, 计算结果正确率后对模型进行评估**【结果】** ChatGPT-4o 与 ERNIE BOT 分别达到了 68.68% 与 66.89% 的正确率, 高于人类考生平均正确率 63.12% **【局限】** 本研究仅针对几个主流模型进行了测试**【结论】** 大语言模型可以理解公务员考试内容, 并具备作为优化考试内容工具的潜力

1 引言

最近, 大型语言模型引起了学术界和工业领域的极大兴趣。正如现有工作所证明的那样, 大型语言模型的出色表现让他们有望成为这个时代的通用人工智能。它们的应用甚至超出了传统的 NLP 领域, 因为它们已被整合到不同领域, 以应对现实世界的挑战[1-3]。由于其在处理一般自然语言任务和特定领域任务等不同应用程序方面的出色性能, 大型语言模型越来越多地被具有关键信息需求的个人(例如学生或患者)使用[4]。

自 1999 年以来中国高等教育毛入学率持续上升, 至 2020 年达到 54.4%, 导致高校毕业生人数从 2000 年的 94 万增至 2022 年的约 1076 万, 创历史新高。然而, 就业市场供需失衡, 使得公务员考试成为众多毕业生的重要就业途径 [5][6]。中国公务员考试包括《行政职业能力测试》和《申论》两部分。《行政职业能力测试》侧重逻辑推理、数据分析、语言理解等综合能力, 《申论》则考查写作、政策分析及综合应对能力。

本研究系统评估了 ChatGPT-4o 等国内外主流大语言模型在 2022-2024 年中国公务员考试中的表现, 揭示其在解决高难度、文化特定任务中的潜力与局限性。同时, 对比各模型在不同领域的表现, 探讨训练数据与模型设计对成绩的影响。

2 相关研究

过去十年中, 自然语言处理领域的最前沿技术取得了显著进展。实际上, NLP 领域的前沿工作经历快速转变, 传统的 NLP 方法已经被基于神经网络的方法所取代。虽然神经技术有着悠久的历史, 但当前的建模方法通常可以追溯

来自 CPU 训练的浅层嵌入式网络架构到现在优化为专用分布式 GPU/TPU 基础设施上运行的基于的神经网络架构[7][8]。在 NLP 技术的加持下大语言模型应运而生，其提供了一个很好的文字数据转换方式，使复杂的文字成为能被神经网络理解的方式。

最近的研究显示，大语言模型可以在医学、法律等领域考试中取得近及格的分數，Bommarito MJ 等人证明 ChatGPT-4 可以通过律师资格考试 [9]；Ioannis Skalidis 等人证明 ChatGPT 能在欧洲核心心脏病考试取得及格成绩 [10]；Panagiotis Tsoutsanis, Aristotelis Tsoutsanis 等人证实 Bing Chat 在作答多专业招聘评估考试中的成绩超过了人类考生平均成绩 [11]。这些研究都清楚地表明了大语言模型在文本处理和理解方面的熟练程度，引发了我们对其在中国公务员考试等高风险标准化考试中的适用性的思考。如 [12] 中所述，OpenAI 的模型是在精选的 CommonCrawl 数据和高质量参考数据的组合上进行训练的，其中数据集不含有中国公务员考试试题。考虑到公考的复杂性与特色性，这是验证大语言模型能力的良好试题[13]。同时，目前主流的大语言模型能否在中国公务员考试中获得成功仍是个未知的问题。

3 方法

3.1 数据集

1995 年至今，公务员考试笔试部分基本未改，还是考《行测》和《申论》两科，只是在内容上按行政级别做些深浅程度的区分[13]。本研究准备了 2022-2024 年的中国国家公务员考试的完整试题。这些试题可以通过公务员考试辅导网站*获得。其中国家公务员行政能力测试包含 130 道单选题，按考察能力不同，分为五个部分，分别是常识判断，言语理解与表达，数量关系，判断推理和数据分析，如表 1 所示；申论部分为给定材料及要求进行写作，有五则材料，共五小题。

3.2 提问范式

由于已证实了提示词对生成式大型语言模型输出具有显著影响 [14]，本研究统一了数据集的输入格式。输入格式统一为引导输入词+原始问题+原始选项，其中问题与选项间以换行符分割，如图 1 所示。

图 1 提问范式

Fig1 question paradigm

* <https://www.fenbi.com/>

示例提问：

请回答下面的单选题。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

根据题目定义，同种数罪指的是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。

ChatGPT-4o回答：

我们来逐一分析各个选项：

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。这属于两起不同的犯罪。甲的行为涉及 过失致人死亡（幼女误饮毒药导致身亡）和自杀（甲自杀）。虽然甲的行为导致了多起死亡事件，但性质上并不属于同种数罪，因为甲的犯罪行为涉及不同的罪名（过失致人死亡和自杀）。

不合同种数罪。

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。乙的行为涉及妨害公务和故意伤害，虽然都是暴力犯罪，但属于两种不同罪名，因此不能算作同种数罪。

不合同种数罪。

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。丙的行为分别构成了 盗窃罪 和 抢劫罪，这两种犯罪的性质和构成不同，因此不属于同种数罪。

不合同种数罪。

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。丁的行为都涉及 故意杀人，尽管是不同的受害人，但两次犯罪行为都符合故意杀人的基本犯罪构成，因此构成同种数罪。

符合同种数罪。

因此，正确答案是 D。

表 1 国家公务员考试行测行政执法卷试卷组成

Table1 National Civil Service Examination Administration and Law Enforcement Paper Paper Composition

CNCSE Component	Total CNCSE Points	Questions	Times
General Knowledge	10 Points	20	2 Hours
Words Understanding and Expression	30 Points	40	
Quantitative Relations	15 Points	15	
Sequitur	25 Points	35	
Data Analyzed	20 Points	20	

3.3 测试步骤

如上文所述，本研究输入采用统一范式，将标准化数据集输入至网页版大语言模型中，每次输入仅输入一小题。模型给出答案后，记录下输出内容，切换至新窗口，清空上下文内容后，继续输入下一题。

3.4 数据分析

在得到大模型作答的全部试题答案后，对大模型生成的答案进行评分。评分过程依据标准答案进行，通过比较大模型的输出与标准答案的一致性来确定正确与否。每个大模型的总得分由其正确回答的题目数量累积而成。随后，根据试题的原始题型进行分部评分。具体而言，将试题按照不同的部分（如常识判断、言语理解等）进行分类，分别计算每个大模型在各个题型中的得分。这一过程有助于深入分析大模型在不同类型题目上的表现差异。接下来，计算每个大模型在各个题型中的百分比正确率。百分比正确率的计算公式为：

$$\text{正确率} = \left(\frac{\text{正确回答的题目数量}}{\text{该题型的总题目数量}} \right) \times 100\%$$

最后，对比各大模型在每个题型中的正确率，得出各模型的评分百分比。

4 结果与分析

4.1 行政能力测试结果分析

(1) 常识判断

表 2 中国公务员考试《行政职业能力测试》常识判断题型模型正确率

Table2 Correctness of General Knowledge Judgment Question Models for the Test of Administrative Aptitude of the Chinese Civil Service Exam

题型分类 (对 应题数)	ChatGPT-4o	Gemini	Copilot	Coze	ERNIE BOT	人类考生
政治常识	5/5	2/5	2/5	3/5	5/5	52.68%
经济常识	1/1	1/1	1/1	1/1	1/1	34%
科技常识	1/3	1/3	1/3	1/3	2/3	41.53%

* <https://www.fenbi.com/>

人文常识	1/2	0/2	1/2	2/2	2/2	47.45%
地理国情	2/2	2/2	1/2	2/2	2/2	38.40%
法律常识	6/7	6/7	2/7	7/7	6/7	38.89%
总体正确率	89.47%	57.89%	57.89%	72.5%	84.21%	40.30%

本部分测试结果如上图表 2 所示。

常识判断中 ChatGPT-4o、Gemini、Copilot 和 ERNIE Bot 的总体正确率分别是 89.47%、57.89%等。

ChatGPT-4o: 在 20 个题目中总得分为 16 分，表现较为出色。特别在法律领域得到了 6/7 的高分，显示其对法律相关知识有较强的理解能力。**Gemini:** 总得分 12/20，表现普通。在技术类知识题目中得分仅为 2/5，显示在技术知识的掌握上有待提高。**Copilot:** 总得分为 8/20。只在某些领域如地理和法律有极为勉强过关的表现，但在技术和科学知识点上得分较低。总的来说表现最差。**Coze:** 总得分 16/20，与 ChatGPT-4o 并列。它在法律领域表现尤为出色，获得了满分 7/7。**ERNIE Bot:** 总得分最高，达到 18/20。在多个领域如技术、人文、法律均有很强的表现，显示了其广泛的知识覆盖和理解深度。

在综合比较这些模型的得分后，可以发现各模型在处理通识知识题目上的能力存在显著差异：

ERNIE Bot 表现最为优异，总得分 18/20，几乎在所有领域都表现出了高水平的理解和回答能力，尤其是在技术和法律领域。

ChatGPT-4o 和 Coze 同时得到 16/20 的较高分，显示出它们在多个知识领域的强劲表现。尤其是 Coze 在法律类题目上获得满分，显示了其在该领域的专业性和深度。

Gemini 和 Copilot 的总得分相对较低，分别为 12/20 和 8/20。这可能指示它们在某些领域的知识掌握不够深入，或者解答策略需要进一步优化。

这次评估明确显示了各个 AI 模型在通识知识处理方面的表现和特长。对于用户来说，根据需求选择最合适的模型非常关键。例如：

如果用户需要在法律或技术类问题上获取准确的回答，ERNIE Bot 显然是最佳选择。

对于需要在多个知识领域内获得较好表现的场景，ChatGPT-4o 和 Coze 可能是更合适的选项。对于对成本和性能有特定要求的应用场景，了解各模型的弱点同样重要，比如 Gemini 和 Copilot 在某些知识领域的表现不佳，可能需要进一步的调整和训练。

(2) 言语理解与表达

在行测考试中的言语理解与表达的这一类题目，单个题目占分 0.8 分一共有 40 道题目。在所有题型里属于占比较大的一种。主要考察的是在中文语境中的理解和填空，其中包括词语成语，和句子。其中每个板块分别有 20 道题目 [14]。

在词语成语方面，大多数考题都是基于成词语的填空来考察的。考生需要关注上下文的具体含义以及语义关系来选出最合适的答案。填空的类型有：单成语，单词语+单成语，双成语，双词语，三成语，单词语+双成语，双词语+单成语七种类型。考察范围全面详细，能很好的反应出考生的语义理解以及词汇运用能力。

在句子方面有填空，分析和排序三种分题型。首先是填空题考生需要理解句子大意之后再对应题目中的空位理解语境，掌握上下文关系，从选项句中找到相匹配的填入空格使上下文连贯顺畅。对于分析题型更多考察的是考生的文段理解和概括能力，有部分题型要求选择文段大意，以及文段没有提及那些部分。最后是句段排序题，题中有 6 个被打乱顺序的句子，考生需要找到内在关系，并在句子连接符合上下文的连贯的规则。对考生的文段结构有较全面的考察。

实验对五种大语言模型进行了 2022-2024 年的行测题目测试。实验步骤：1. 将题目以文字的形式输入到大模型中，2. 记录其得出的结果和对应解析并加以分析，3. 统计结果分析数据。五种大模型的实验过程有些许区别，主要是集中在实验步骤 1 中。对应有些大模型没有图片识别功能如 Coze 和 ERNIE BOT。因此改变了题目的输入方式，采用文字的形式在同一个话题中输入并继续实验剩余步骤。在对话预设方面我们只告诉了模型“你将要参加一个测试请给出你的答案以及对应解释”。此目的旨在保证模型能够对题目有充分理解，没有对测试结果产生影响。

模型测试结果如表 3，表中展示了 5 种大模型在言语理解与表达的分题型中回答表现。其中最后一行是每个大模型在言语理解与表达题型中的总正确率，以及标注了人类考生的平均成绩以供参考。

可以看到五种模型大多有比较好的回答结果，其中 ERNIE BOT 和 ChatGPT-4o 取得了较高的正确率，分别是 87.8% 和 75.0%。说明这两个大语言模型在言语理解与表达的部分有充足的理解并且可以很好的解决这一类问题。再看其他三种大语言模型正确率从高到低是 Copilot, Coze, Gemini, 分别是 72.6%, 72.5%, 70.6%。虽然最低的 Gemini 正确率也有 70.6% 的正确率，但其还是高于人类平均测试水平，说明其有一定的语义理解能力的。表 3 中可以看到各模型在 2022-2024 年卷中的不同分题型上的表现。

表 3 中国公务员考试《行政职业能力测试》言语理解与表达题型模型正确率
Table 3 Correctness of Verbal Comprehension and Expression Question Models for the Chinese Civil Service Exam Test of Administrative Aptitude (TOCA)

题型分类 (对应题数)	ChatGPT-4o	Gemini	Copilot	Coze	ERNIE BOT
单成语	1/3	2/3	1/3	2/3	2/3
单成+单词	4/5	4/5	2/5	5/5	3/5
双成语	5/5	5/5	4/5	5/5	4/5
双词语	1/1	1/1	1/1	1/1	1/1
三成语	0/1	0/1	0/1	0/1	1/1
一词+二成	2/2	1/2	1/2	2/2	1/2

* <https://www.fenbi.com/>

二词+一成	2/3	2/3	2/3	3/3	2/3
句子补全	3/3	3/3	3/3	3/3	3/3
中心主旨	3/5	4/5	5/5	5/5	5/5
句子排序	3/3	1/3	1/3	2/3	3/3
语义检测	9/9	9/9	7/9	8/9	9/9
总体正确率	75.0%	72.6%	70.6%	72.5%	87.8%

*人类考生总体平均正确率：69.2%

总得来说，大模型在回答语义理解问题上有一定的能力。最主要的区别是模型是否会进行语境的分析，如 Copilot 在回答问题使无法经进行良好的语义理解，无法掌握当下语言环境的特征和重点，语义理解能力差。反观 ChatGPT-4o，在回答问题的时候有明显的语义理解过程，将词语的含义带入文章，判断上下文是否合适。符合语义理解的正确做题思路，在测试中表现出了语义理解的逻辑能力。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.训练集大小 2.训练模型的不同方向 3.训练程度不同。因为这几点的不同，导致不同模型在语义理解这种题型上产生了差距。

(3) 数量关系

表 4 数量关系类别目各小题模型正确率

Table 4 Correctness of the model for each subtopic of the quantitative relations category item

题型分类 (对应题数)	ChatGPT-4o	Gemini	Copilot	Coze	ERNIE BOT
22 年正确率	50.00%	20.00%	80.00%	50.00%	40.00%
23 年正确率	60.00%	40.00%	50.00%	70.00%	30.00%
24 年正确率	70.00%	30.00%	50.00%	60.00%	50.00%
平均正确率	60.00%	30.00%	60.00%	60.00%	40.00%

*人类考生总体平均正确率 38.73%

如上图表 4 所示。

根据数量关系内容分类的结果：

在不同领域问题上的表现差异揭示了 AI 模型在处理特定类型问题时的能力。人类测试者通常会根据问题的性质表现出不同的水平，AI 模型同样如此。表 3 展示了五个模型——文心一言、Copilot、ChatGPT-4o、Coze 和 Gemini——在三年行测中的不同正确率。结果（表 4）表明，这些模型在不同年份问题上的表现存在显著差异。根据其他对于大模型们的研究不难看出，大模型在数学推理方面的熟练程度仍然需要提高

ChatGPT-4o 在总体上表现最为出色，在 23 年和 24 年都是各模型中表现最好的，这很可能是因为此模型的数学计算能力相比其他模型而言最为优秀。同时，从众模型的错误原因中，很容易发现它们大部分在读取题意上出现了错误，这也反映出 ChatGPT-4o 或许拥有更加优异的题意理解能力。

* <https://www.fenbi.com/>

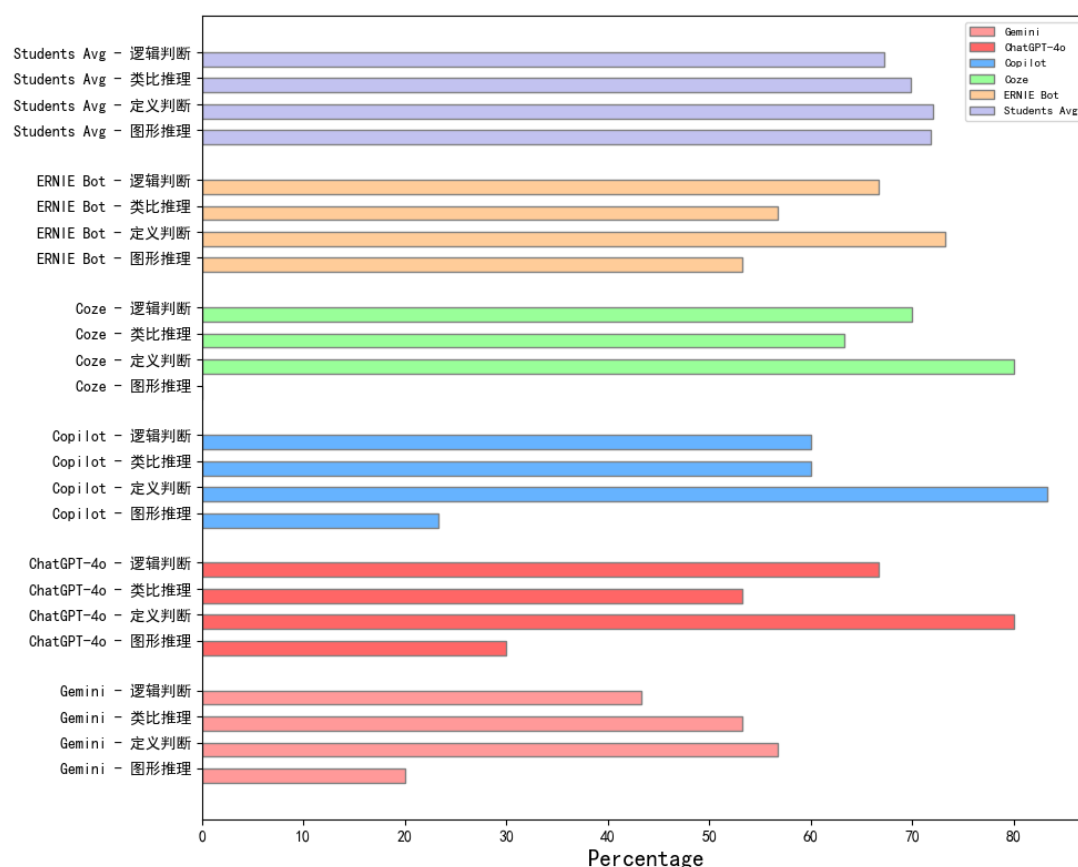
值得注意的是，在 2022 这一年中，所有模型的表现都较差，准确率均为 0%。这可能反映了这年题目在这些模型中的难度较大，或是当前模型在处理此类任务时仍存在不足，大模型们在数学推理上的能力可能尤为不稳定。总体而言，大型语言模型在数学推理方面表现出巨大的潜力，并表现出持续改进的趋势，但仍面临许多挑战和限制，需要更深入的研究和优化。

综合言之，ChatGPT-4o 在这些模型当中水平最高，在处理数量关系类的数学计算题目中更具备优势。

(4) 判断推理

图 2 判断推理部分各模型正确率数据

Fig 2 Data on the percentage of correctness of each model in the judgmental reasoning section



在本轮测试中，表现最佳的是 ERNIE Bot 模型，总正确率为 62.5%。ERNIE Bot 不仅取得了最高的正率，且所有模型中正确率的方差最小，体现了其在不同任务下稳定的性能表现。在具体任务的表现上，ERNIE Bot 表现出一定优势，多项任务位列第一，在图形推理与类比推理任务中，分别获得了 53.3%和 56.7%的正确率。此外，在定义判断任务中，ERNIE Bot 达到了 73.3%的高正确率，在逻辑判断任务中，其正确率为 67.3%。

排名第二的是 ChatGPT-4o 模型，总正确率为 57.5%。在具体任务上，ChatGPT-4o 在定义判断中取得了 80%的高正确率，在类比推理和逻辑判断任务

中，分别取得了 53.3%和 66.7%的正确率。然而，在图形推理任务中，其正确率仅为 30%，表现相对较弱。

紧随其后的是 Copilot 模型，总正确率为 56.7%，与 ChatGPT-4o 模型相差不到 1%。在具体任务表现上，Copilot 在定义判断任务中取得了单项最高的 83.3%正确率，在类比推理和逻辑判断任务中，均获得了 60%的正确率；在图形推理任务中，其正确率稍低，为 23.3%。

Coze 模型以 53.3%的总正确率位列第四。在具体任务中，Coze 在定义判断任务中获得了 80%的正确率，在类比推理任务中取得了 63.3%的正确率，而在逻辑判断任务中，以 70%的正确率位列单项第一。然而，在图形推理任务中，Coze 的正确率为 0%，出现了多题无法正常识别和回答的情况。

本轮测试中，平均正确率最低的是 Gemini 模型，仅达到 43.3%。在图形推理、定义判断、类比推理和逻辑判断四项任务中，其正确率分别为 20%、56.7%、55.3%和 43.3%。

人类考生在判断推理专项下获得了 69.35%的总正确率，领先于所有大型语言模型。同时，成绩总览图 2 显示，人类考生在四个子项目中的正确率较为均衡，方差较小，每个子项目的正确率均超过 60%，其中图形推理任务的正确率最高，为 72.5%，逻辑判断任务的正确率最低，为 66.3%。相较于大型语言模型，人类考生的考试表现更加均衡，各子项目的正确率均在 70%上下浮动。

本轮测试结果表明，即使是目前最先进的多个大型语言模型，在处理公务员行测考试的判断推理试题时，其表现仍不尽如人意，主要表现为图形题识别不准确、推理逻辑混乱，答题正确率低于人类考生的平均水平。因此，不建议将其作为判断推理部分的辅助工具使用。

(5) 数据分析

表 5 2022-2024 年数据分析部分各模型得分情况

Table 5 Scores for 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%
Copilot	60%	45%	45%	50%
ERNIE BOT	35%	20%	35%	30%
Gemini	30%	30%	20%	26.5%
Coze	15%	15%	20%	16.5%
Human	73.05%	73%	72.7%	72.9%

本轮测试结果显示（表 5），ChatGPT-4o 的表现最佳，平均准确率 61.5%；Copilot 以 50%的准确率排名第二；其他模型如 ERNIE Bot、Gemini 和 Coze 的表现相对逊色，分别获得了 30%、26.5%和 16.5%的正确率。所有模型的平均准确率均不及人类考生，且这些模型的表现差异揭示了当前人工智能模型在处理复杂推理和数据分析任务时的能力差异。

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 6 所示。

表 6 各模型 2022-2024 年行测平均正确率

Table 6 The average accuracy rate of each model in the Administrative Competency Test (ACT) from 2022 to 2024

模型	Gemini	Coze	Copilot	ERNIE BOT	ChatGPT-4o	人类考生
平均正确率	50.95%	57.02%	54.18%	66.89%	68.68%	63.12%

4.2 申论

申论部分测试结果如表 7 所示,我们选取在行测环节中取得最高正确率的 ChatGPT-4o 模型来进行测试。在申论测试中,ChatGPT-4o 的平均正确率已显著超过了人类考生的平均正确率,但仍与最优秀的人类考生有明显差距。这表明,ChatGPT-4o 在文字综合理解与写作能力方面,其表现已具备较高的实际应用水平。

表 7 中国公务员考试申论部分的 ChatGPT-4o 得分

Table 7 ChatGPT-4o scores in the essay section of the Chinese Civil Service Examination

年份	ChatGPT-4o	人类考生(平均/最高)
2024	65.5%	30.4/86
2023	60%	28.36/85
2022	58%	32.4/90
平均正确率	61.17%	30.38%/87%

4.3 结果比较与案例分析

在众多大语言模型做出的回答中,总结了部分大语言模型的答题特点,并用相关例题进行了举例分析。详见附录。

5 结论

为探索大语言模型在复杂、高风险标准化考试中的应用前景与局限性。本文对大语言模型在中国公务员考试中的能力进行了测评。本次测评首先收集了考题信息,并规定了提问范式,进而将数据按范式形式输入模型中,并记录输出结果,计算正确率后对模型进行评估,并根据部分案例进一步分析模型生成质量。结果显示 ChatGPT-4o 与 ERNIE BOT 两个模型取得了 68.68%与 66.89%的正确率,超过了人类考生平均正确率,具有在中国公务员考试中的应用前景。

在测评过程中,本研究也探索到一些问题。当下的大语言模型普遍在图像识别能力上有所欠缺,在逻辑推理方面能力不足。五种大语言模型在逻辑推理部分的得分依次为: ERNIE Bot 62.5%, ChatGPT-4o 57.5%, Copilot 56.7%, Coze 53.3%, Gemini 43.3%, 而人类考生的平均成绩为 69.35%相对高于各大语言模型。因此,想要在具体领域有良好的大语言模型辅助功能,需要针对对应

领域训练特定的模型,这样才能让大模型帮助特定领域的发展[15][16]。本文也存在一些不足之处,例如,模型的验证仅限于公务员考试公开数据集,未来将收集更多真实且包含其他模态的数据进行测试,本文测评模型数量较少,未来将涉及更多的大语言模型进行测试。

参考文献:

- [1] Jungo Kasai, Yuhei Kasai, Keisuke Sakaguchi, Yutaro Yamada, and Dragomir Radev. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. In ArXiv preprint, 2023. ArXiv:2303.18027, <https://arxiv.org/abs/2303.18027>.
- [2] Nikhil Savale. Quantitative version of weyl's law. In ArXiv preprint, 2023. ArXiv:2303.01407, <https://arxiv.org/abs/2303.01407>.
- [3] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. In ArXiv preprint, 2023. ArXiv:2303.17564, <https://arxiv.org/abs/2303.17564>.
- [4] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. 2024.
- [5] 张仲琪; 高翠欣;. 后疫情时代大学生“考公务员热”现象研究. 2022(Zhang Zhongqi; Gao Cuixin;. Research on the phenomenon of college students' "civil service examination fever" in the post-epidemic era. 2022.)
- [6] 吴奕. 为公务员热降温还需哪几招. 2018(Wu Yi. What are the measures needed to cool down the civil servant fever? 2018.)
- [7] Shum HY Duan N Liu S, Zhou M. 2020 progress in neural nlp: modeling, learning, and reasoning. In Engineering 6, page 275 – 290, 2020.
- [8] Kalita JK. Otter DW, Medina JR. 2020 a survey of the usages of deep learning for natural language processing. In IEEE Trans. Neural Netw. Learn. Syst. 32, page 604 – 624, 2020.
- [9] Pablo Arredondo Daniel Martin Katz Michael James Bommarito, Shang Gao. Gpt-4 passes the bar exam. 2024.
- [10] Wongsakorn Luangphiphat Thabo Mahendiran Olivier Muller Emmanuel Abbe Stephane Fournier Ioannis Skolidis, Aurelien Cagnina. Chatgpt takes on the european exam in core cardiology: an artificial intelligence success story? 2023.
- [11] Aristotelis Tsoutsanis Panagiotis Tsoutsanis. Evaluation of large language model performance on the multi-specialty recruitment assessment (msra) exam. 2024.

- [12] Ryder N Subbiah M Kaplan JD Dhariwal P Neelakantan A Shyam P Sastry G Askell A Agarwal S. Brown T, Mann B. Language models are few-shot learners. 2020.
- [13] Zhang W Liu, Z. A qualitative analysis of chinese higher education students' intentions and influencing factors in using chatgpt: a grounded theory approach. 2024.
- [14] Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. On the relation between sensitivity and accuracy in in-context learning. In arXiv preprint arXiv:2209.07661, 2024. <https://arxiv.org/abs/2209.07661>.
- [15] Thomas Hofmann Abu Sebastian Abbas Rahimi Michael Hersche, Francesco di Stefano. Probabilistic abduction for visual abstract reasoning via learning rules in vector-symbolic architectures. 2024.
- [16] Tianyu Liu Wei Wang Yijia Xiao, Edward Sun. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. 2024