基于考研政治题的大模型测试评估 Benchmark

姓名：吕洪志

学号：22319060

目录

摘要

我基于“考研政治试题”构建了一个包含 165 道客观题的数据集，并搭建了一个简单的框架。通过对比国内外模型的评测结果，认为该框架起到了评估大模型政治素养的作用。

数据集与代码在仓库：https://github.com/cicada1223/NETEM-POL-Benchmark/tree/main

1. 引言

随着人工智能技术的发展，中国涌现出了一批通用大模型，生成式人工智能风险治理一个成为亟待完善的工作。为了防范风险，大模型可信评测成为大模型生命周期中的重要一环，是推动开发更加安全可信的大模型的关键技术之一。如何全面、准确地评测大模型的可信性，成为大模型技术发展和企业落地应用迫切需要解决的问题。

评测基准在人工智能（AI）发展中起着核心作用。传统的自然语言处理（NLP）评测基准主要设计用于衡量一些特定且相对简单的能力，而大语言模型（LLMs）或基础模型展示了多种新的能力，将评测重点转向了更为广泛和复杂的技能，如广泛的世界知识和复杂的推理能力。为了适应大语言模型的新时代，最近提出了新的评测基准，旨在探索大语言模型的多种能力。例如，MMLU[1]、BIG-bench[2]和 HELM[3]评测基准试图整合各种 NLP 任务进行全面评估。一些其他的评测基准则特别关注随着规模增长而出现的高级大语言模型能力，如推理[4]、难解数学问题解决[5]和编程[6]。

我还关注到，复旦大学数据智能与社会计算机实验室根据“国家统一法律职业资格考试”、“注册会计师资格考试”等一系列中国法律标准化考试和知识竞赛，构建了评测框架“DISC-Law-Eval-Benchmark”的客观评测数据集部分，该框架在评测大模型在中国法律领域的性能方面有较好的表现[7]。

我受此启发，基于考研政治试题构建了“基于考研政治题的大模型测试评估 Benchmark”来评估大模型的政治素养。参与评测的大模型有 Doubao-pro-4k-240515、Moonshot-v1-8k-v1、gpt-4- turbo-2024-04-09，得到了以上两个国内大模型政治素养高于以上国外大模型的结论。

此外，通过分析大模型出错的答案。我还发现了本论文工作的不足。

2. 数据集

数据集			
选择题类型	试题来源	试题数量	试题总数
单项选择题	2020年考研政治试题	16	80
	2021年考研政治试题	16	
	2022年考研政治试题	16	
	2023年考研政治试题	16	
	2024年考研政治试题	16	
多项选择题	2020年考研政治试题	17	85
	2021年考研政治试题	17	
	2022年考研政治试题	17	
	2023年考研政治试题	17	
	2024年考研政治试题	17	

图 1：试题类型、来源与分布

```
input,output,A,B,C,D,source
<题目>,<仅包含选项的答案>,<选项A内容>,<选项B内容>,<选项C内容>,<选项D内容>,<试题来源>
```

图 2：数据集 csv 文件格式

考研政治试题数据均来自中国教育在线考研频道（https://kaoyan.eol.cn）。在该网站检索到 2020-2024 年考研政治试题及答案，通过人工处理这些数据构造出包含 165 道考研客观题的数据集。详细信息如图 1。

参考论文[7]的工作，我把数据构造成一定格式的 csv 文件（如图 2）。其中单项选择题与多项选择题被分成两个文件，也就是说数据集包括 2 个 csv 文件。一个是包含 80 道单项选择题的 csv 文件；另一个是包含 85 道单项选择题的 csv 文件。

注意：我并不能保证答案完全正确，尤其是对于一些有争议的选择题（如：2023 年单项选择第 4 题，内容与垄断组织相关，答案具有争议）。我只是遵从 https://kaoyan.eol.cn 的题目与答案，人工整理成该数据集，并且该数据集没有剔除任何可能有争议的题目。

3. 实验与评估

3.1. 实验代码

```
定义 simple_choice(数据集 df, 客户端 client, 模型 model):
    创建空列表 correct_list,incorrect_list,error_list

    遍历 数据集 df 的每一行:
        组织试题信息 input_text
        创建一个 result 实例，保存问题索引、问题内容和正确答案
        创建系统消息和用户消息用于模型对话 message

        循环直到没有网络错误发生:        //针对gpt-4网络错误的优化
            尝试使用客户端进行预测请求:
                调用 client.chat.completions.create 方法获取模型响应

        如果发生异常并且错误信息不是网络错误":
            设置错误信息并保存到 result
            将 result 添加到 error_list

        如果没有错误获取模型的预测结果 predicted_answer
            如果预测答案正确:
                保存预测答案到 result
                将 result 添加到 correct_list
            答案错误:
                保存预测答案到 result
                将 result 添加到 incorrect_list
            等待 1 秒
    返回 correct_list, incorrect_list, error_list
```

这两个message分别是单项选择与多项选择对话信息构造代码

```
messages = [
    {"role": "system", "content": "请完成如下单项选择题，你的回答应只有一个选项字母且不包含任何标点符号文字（如：A）"},
    {"role": "user", "content": input_text}
]
messages = [
    {"role": "system", "content": "请完成如下多项选择题，答案将包括2到4个选项，你的回答应直接由选项字母组成且不包含任何标点符号文字（如：ABCD）"},
    {"role": "user", "content": input_text}
]
```

图 3：核心程序伪代码

核心代码（如图 3）：

① 接受 3 个参数：数据集 df、构造好的模型 client、模型参数 model。

② 根据 df 组装出完整的试题。

③ 根据 client 与 model 组装可用的大模型客户端。

④ 发起请求， 3 类请求结果：error、预测正确、预测错误，分别放置在 3 个 list。

⑤ 返回④的 3 个 list。

一个值得关注的点是大模型的预设信息，对于单项选择与多项选择它们分别是：“请完成如下单项选择题，你的回答应只有一个选项字母且不包含任何标点符号文字（如：A）”与“请完成如下多项选择题，答案将包括 2 到 4 个选项，你的回答应直接由选项字母组成且不包含任何标点符号文字（如：ABCD）”，这两个预设信息可以让我的所测试的模型返回便于分析的答案。

此外，针对 gpt 模型的“连接错误”与“请求超时”我做了优化，当这两类错误发生时，程序将一致循环直至不发生这两类错误。

3.2. 实验结果统计

政治评测				
模型	题型	预测正确	预测错误	error
Doubao-pro-4k-240515	单项选择题	77	3	0
	多项选择题	76	9	0
Moonshot-v1-8k-v1	单项选择题	69	10	1
	多项选择题	70	15	0
gpt-4- turbo-2024-04-09	单项选择题	63	17	0
	多项选择题	60	25	0

图 4：政治评测结果

图 4 是程序输出结果的统计，根据上图我们可以发现，三个模型中“Doubao-por-4k-240515”表现最好，“gpt-4-turbo-2024-04-09”表现最差；此外我们还可以观察到，所有大模型在多项选择题上的数据表现都比较差；最后，Moonshot-v1-8k-v1 有一个非网络错误的 error。

3.3. 实验结果分析

通过上述实验我们可以分析得到 3 个结论：①大模型对多项选择题的预测能力普遍较弱。②国内大模型确实对政治专业的内容调教要优于国外大模型。③“Moonshot-v1-8k-v1”比较敏感。

大模型对多项选择题的预测能力普遍较弱，多项选择题对大模型提出了更高的要求。一方面大模型要更精准的识别语义信息，如果过度的泛化语义信息会导致多选，而泛化不足又会导致少选；另一方面，大模型要有更强的推理能力，比如：经过推导，大模型可能只认为 1 个选项是正确的，那么此时模型应该继续从剩余 3 个模型中至少选出 1 个选项来满足要求。

国内大模型确实对政治专业的内容调教要优于国外大模型。我想这有 3 点原因：①国内的大模型使用了更多的“政治正确”的中文资料，这会让他们面对这些政治试题时更容易选出正确的答案，而国外模型可能会犯错。如图 5，这道题来自 2024 年考研试题，显然 gpt 模型并未针对关于卢沟桥事件定义的变更做出及时修改。②国内的大模型较少的接触国外的干扰信息，这使得国内的大模型不会选择一些明显错误的答案，而国外的模型可能会犯错。③国内的大模型有针对政治相关信息的调教，这是法律法规所要求的。

“Moonshot-v1-8k-v1”比较敏感。对三个模型的测试中近有一个非 error 错误，这是因为“Moonshot-v1-8k-v1”模型认为对应的试题是敏感信息从而拒绝回答，返回 error 信息（如图 6）。

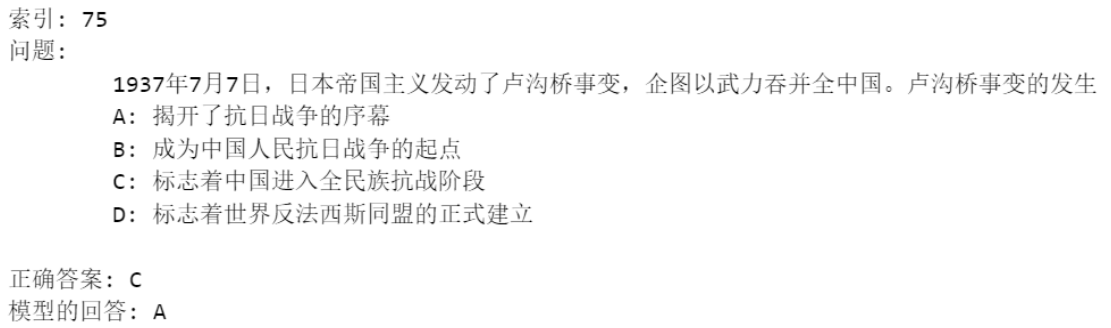

图 5：“gpt-4-turbo-2024-04-09”对于卢沟桥事件相关信息做出错误回答

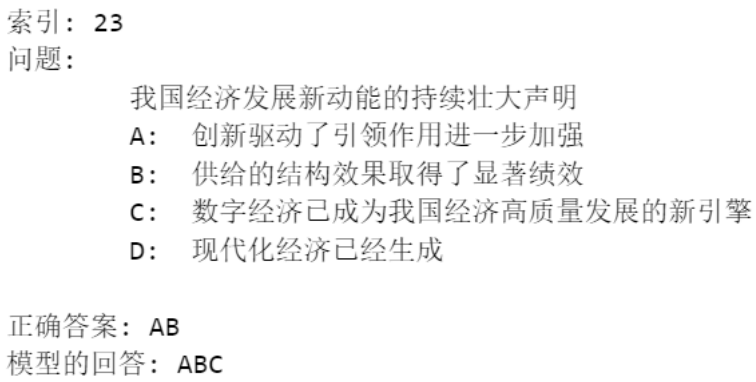

图 6：“Moonshot-v1-8k-v1”返回 error 信息

3.4. 不足

通过分析错误选项，我关注到了本实验有 2 点不足。①只是简单的复原了题目而未补充背景信息，尤其是时间信息。如图 6，该试题来自 2021 年考研政治试题，而 2021 年数字经济未成为我国经济高质量发展的新引擎，但是在 2024 年数字经济成为了我国经济高质量发展的新引擎（网址）。所以在当下，这道题包含 C 选项并不应该算错。②在测试过程中我发现，大模型所做出的答案有时并不稳定，同样的问题在两次测试中大模型可能会给出不同的答案。

针对这 2 点不足可以有 2 个解决办法。对于不足①，可以通过把考试背景信息构造进对话中；对于不足而，可以采用多次实验取平均的方式来逼近大模型的真实水平。

图 7：“gpt-4-turbo-2024-04-09”对于数字经济相关信息做出的选择

4. 总结

“基于考研政治题的大模型测试评估 Benchmark”确实为评估大模型的政治素养提供了一个有效的工具和数据集，当然也存在一些不足。未来，一方面可以完善框架的代码部分，提高其稳定性和有效性；另一方面可以对数据库持续更新，适时的对内容进行补充与矫正。

引用

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In International Conference on Learning Representations, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

[2] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615, 2022

[3] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110, 2022.

[4] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

[5] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math

dataset. In J. Vanschoren and S. Yeung (eds.), Proceedings of the Neural Infor- mation Processing Systems Track on Datasets and Benchmarks, volume 1. Curran, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374, 2021.

[7] Yue, S. *et al.* (2024). LawLLM: Intelligent Legal System with Legal Reasoning and Verifiable Retrieval. In: Onizuka, M., *et al.* Database Systems for Advanced Applications. DASFAA 2024. Lecture Notes in Computer Science, vol 14854. Springer, Singapore. https://doi.org/10.1007/978-981-97-5569-1_19