计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241000129

大模型交叉测评方法研究

梁秉豪　张传刚　袁明明
浪潮通信信息系统有限公司　济南 250013
(liangbinghao@inspur.com)

摘　要　随着ChatGPT的出现,大模型已经成为全球科技竞争的新赛道,并开始广泛应用于生产和生活的各个环节。国内众多科技公司纷纷投入到大模型研发和开源工作中。大模型应用场景不断拓展,可供下载或调用的预训练大模型类型和数量越来越多,用户对于大模型测评的需求逐渐增加。目前面向大模型测评还未形成标准化的方法,业界主要通过第三方机构提供的测评榜单对大模型能力进行横向对比。大模型在特定应用场景下的实际效果仍缺少有效的测评手段。文章针对预训练大模型在垂直行业场景下的应用效果测评,特别是面向开放性问题的回答能力进行研究,提出了一套交叉测评方法,并对其可靠性和鲁棒性进行了实验验证。实验结果表明,所提交叉测评方法测评结果与官方给出结果一致性较高,说明该方法具有较强的可靠性。所提方法有效提高了大模型测评结果的客观性和便捷性,有助于用户在个性化场景中快速完成大模型的横向对比和选型。

关键词:测评方法;交叉测评;开放性问题;待测评大模型;裁判员大模型
中图分类号　TP311

Research on Cross-Evaluation Method of Large Model

LIANG Binghao,ZHANG Chuangang and YUAN Mingming
Inspur Communication Information System Co.,Ltd.,Jinan 250013,China

Abstract　With the emergence of ChatGPT,large model have become a new track for global technology competition,and have begun to be widely used in all aspects of production and life. Many domestic technology companies have invested in large model research and development and open source work. As the application scenarios of large model continue to expand,there are more and more types and quantities of pre-trained large model that can be downloaded or invoked,and users' demand for large model evaluation is gradually increasing. At present,there is no standardized method for the evaluation of large model,and the industry mainly compares the capability of large models through the evaluation lists provided by third-party institutions. There is still a lack of effective measurement methods for the actual effect of large models in specific application scenarios. In this paper,a cross evaluation method is proposed to evaluate the application effect of the pre-trained large model in the vertical industry scenario,especially the answering ability of open questions,and its reliability and robustness are verified by experiments. The cross-evaluation method proposed in this paper has a high consistency with the official results,indicating that the method has a strong reliability. This method effectively improves the objectivity and convenience of large model evaluation,and helps users to quickly complete the horizontal comparison and selection of large models in personalized scenes.

Keywords　Evaluation method,Cross evaluation,Open-ended question,Candidate large model,Judge large model

1　引言

2023年以来,以ChatGPT为代表的AIGC技术快速发展,人工智能为数字经济的发展带来新的技术红利。其中大模型展现出惊人的能力,使得其不管是在学术界还是在工业界都备受青睐。随着大模型不断走向商用,并在日常工作和生活中被广泛使用,如何对其进行有效的性能、安全性等方面的测评变得至关重要。目前围绕大模型测评的研究较多,但尚未有统一且有效的方法对大模型进行客观、公正的系统性测评。对大模型进行测评可以量化对比不同大模型的优劣势,然而大模型的部分能力来源于对数据集记忆,部分离线测评方法和评估基准被指出可能存在题库泄密风险,导致测评结果客观性存疑。如何设计一套有效的大模型测评方法,对大模型技术的发展,以及用户的技术选型,都有着重要的意义。

现有大模型测评方法主要分为两类:标准题库测评和专家经验测评。标准题库测评方法主要利用大规模题库进行测评,通过对比题库标准答案和大模型返回答案完成打分,测试范围局限于题库中的内容,且对于开放性问题的回答结果难以进行评估。专家经验测评方法主要通过人类专家的主观判断,对大模型生成的答案进行打分。该方法难以实现自动化测评,且不同专家打分可能存在差异,测评结果客观性和可重复性不足。

2　相关工作

在标准题库测评方面,针对大语言模型的基准测试主要包括自然语言理解和自然语言生成两大类任务[1]。对于通用

基金项目:泰山产业领军人才项目(tscx202312006);山东省博士后创新项目(SDCX-ZG-202400307)
This work was supported by the Taishan Industrial Leading Talent Project(tscx202312006) and Shandong Postdoctoral Innovation Project(SDCX-ZG-202400307).
通信作者:张传刚(zhangchg@inspur.com)

大模型综合性能已有较多基准测试方案。纽约大学和华盛顿大学等机构开发的 GLUE[2] 一共包含了 9 类自然语言理解任务，涵盖了语义相似度和情感分析等基础任务。斯坦福大学等机构开发的 SuperGLUE[3] 参考 GLUE 的基本设计，对任务形式进行了丰富，提供了推理和问答等测评形式。微软开发的 AGIEval[4] 包含了多种官方入学考试、资格考试等基准数据，对大模型认知、知识理解和逻辑推理等能力进行全方位的测评。MMLU[5] 覆盖了 57 类任务，其中包括数学、历史、计算机科学和法律等领域的知识，主要通过选择题的方式进行测评。Google，OpenAI 等 100 多家机构联合开发的 BIG-Bench[6]，一共支持 200 多类测评任务。C-Eval[7] 提供了涵盖 52 个不同学科的多项选择题测试集。MMCU[8] 主要覆盖了医疗、法律、心理学和教育等领域测试，包含了一万多个单项和多项选择题。此外，目前还有面向专业场景的基准测试集，如面向金融领域的 FLUE[9]，FineEval[10] 和 CFBenchmark[11] 等，面向法律领域的 LawBench[12] 等，面向医疗领域的 MedBench[13] 等。

在专家经验测评方面，2023 年 5 月，国际开放研究组织 LMSYS Or 推出的 Chatbot Arena 平台通过用户直接参与投票和盲测方式对大模型进行客观测评，完成大模型性能的两两对比。2024 年 5 月，由上海人工智能实验室开源的大模型测评平台，提供了司南大模型竞技场（CompassArena），采用专业用户投票的方式，对多个大模型匿名返回的问题回答进行横向对比和投票，此类方法客观性较强。2024 年 9 月，中国信息通信研究院依托工信部大模型公共服务平台发布了“大模型应用选型竞技赛场”，通过匿名对战和自选对战的方式对多个大模型的回答结果进行对比，支持用户选择较优模型并录入选择的依据。此外，SuperCLUE 和阿里云审明师等产品通过引入超级模型作为裁判进行大模型测评，有效提升了开放性问题下的测评效率。

本文主要研究大模型在自然语言生成任务中的测评方法，将其划分为主观测评和客观测评两大类（如表 1 所列），其中主观测评分为逻辑推理、阅读理解和开放问答，客观测评分为文本分类、文本相似度和判断选择题。针对客观测评任务，通过标准题库测评方式可以快速得到测评结果。针对主观测评任务，目前主要通过专家经验测评方式，通过大模型匿名竞技的方式，借助专家投票得到测评结果。然而现有的测评基准还存在以下几个问题：

（1）测评结果权威性不足：通过题库方式对大模型能力进行测评，覆盖范围较为有限，同时容易因为题库泄露导致测评结果的客观性和参考价值存疑。裁判员大模型在中文场景下的特定细分领域效果难以保证。

（2）测评场景覆盖度有限：目前基准测评数据集的内容较为固定，大多为通识性问题，无法覆盖特定行业场景或特定业务需求。此外，对于缺少标准答案的开放性问题，现有技术难以较好完成测评。

（3）测评过程工作量较大：建立标准题库工作量较大，数据采集、数据清理和数据标注等过程需要大量技术和业务专家参与。此外，通过专家经验投票的方式，需要逐个对大模型回答结果进行对比，效率较低。

（4）测评数据安全难以保证：通过超级模型（如 ChatGPT-4）作为裁判员大模型进行测评时，容易出现数据泄露等安全问题。

针对大模型测评现状和存在的问题，本文设计了一套预训练大模型的交叉测评方法，主要面向主观测评，特别是开放问答场景，通过提示词设计和循环迭代方式以及划分待测评大模型（CLM）和裁判员大模型（JLM），引导大模型进行交叉测评，无需人工干预便可获得大模型测评结果。基于该方法设计并研发了一套面向预训练大模型的交叉测评系统，通过实验数据表明该方法测评效果较好。

表 1 大语言模型测评基准

Table 1 Evaluation benchmark for large language models

类型	任务类型	基准数据集
主观测评	逻辑推理	GLUE，SuperGLUE，AGIEval，BIG-Bench，Chatbot Arena，CompassArena
	阅读理解	SuperGLUE，AGIEval，BIG-Bench，Chatbot Arena，CompassArena
	开放问答	BIG-Bench，Chatbot Arena，CompassArena，CFBenchmark，MedBench
客观测评	文本分类	GLUE，BIG-Bench
	文本相似度	GLUE，SuperGLUE，BIG-Bench
	判断，选择题	GLUE，AGIEval，MMLU，BIG-Bench，C-Eval，FineEval，CFBenchmark，MedBench

3 预训练大模型交叉测评

3.1 系统总体架构

本文所设计的预训练大模型交叉测评系统整体架构如图 1 所示，主要包括测评问题库、大模型接口模块、推理服务模块、模型测评模块和综合评分模块五大模块。

（1）测评问题库包含测试所需的数据集，主要面向语言类大模型在垂直行业应用中实际需要解决的各类问题，包括了各类场景下的开放性问题，用于对比各类大模型在开放性问题中的表现。

（2）大模型接口模块主要用于适配常用大模型推理和聊天等接口，通过 API 接口调用的方式提供大模型基础能力，根据输入的测评问题，输出对应答案。

（3）推理服务模块通过读取测评问题库中的问题，然后调用待测评大模型接口，对待测评大模型回答结果进行存储和管理。

（4）模型测评模块将各个待测评大模型的回答整理成问答对，通过提示词工程调用裁判员大模型，对待测评大模型的回答结果进行交叉评分。

（5）综合评分模块对大模型交叉评分结果进行加权求和，逐个计算大模型测评得分，根据各个大模型测评得分更新其在下一轮迭代过程中所给出打分的权重，当达到收敛条件时输出最终得分。

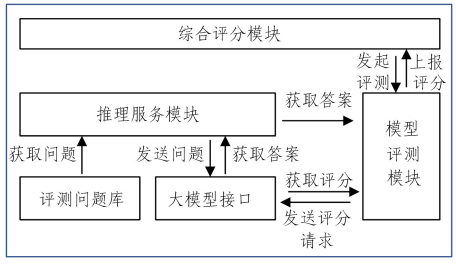

图 1 系统整体架构

Fig. 1 Architecture of the system

3.2 交叉测评流程设计

本文所设计的预训练大模型交叉测评如图 2 所示，主要包括以下步骤：

(1)从测评问题库中随机抽取 k 个问题，形成基准的测评问题集 $Q=\{Q_1,Q_2,\cdots,Q_k\}$；

(2)推理服务模块循环调用 n 个待测评大模型$m^c=\{m_1^c,m_2^c,\cdots,m_n^c\}$的接口，获取每个待测评大模型对测评问题集的答案 $a=\{a_1,a_2,\cdots,a_n\}$；

(3)通过提示词工程引导裁判员大模型$m^j=\{m_1^j,m_2^j,\cdots,m_n^j\}$对待测评大模型进行评分，得到大模型评分矩阵 $\boldsymbol{S}$(本文中裁判员大模型集合与待测评大模型集合所包含的大模型相同)；

(4)根据裁判员大模型权重w_i(表示大模型m_i^j评分的可靠性)对待测评大模型的所有评分进行加权求和，得到综合得分 $s=\{s_1,s_2,\cdots,s_n\}$；

(5)判断是否满足终止条件，若满足，则输出最终评分 s，否则更新模型权重 $w=\{w_1,w_2,\cdots,w_n\}$，然后跳转至步骤 4 重新计算综合得分。

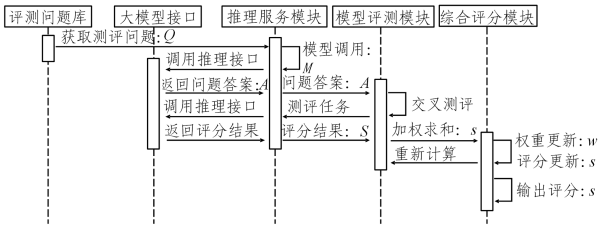

图 2 交叉测评流程图

Fig. 2 Flowchart of cross-evaluation

3.3 大模型回答生成和评分生成

3.3.1 待测评大模型回答生成

本文对待测评大模型的提示词模版进行了设计和验证，发现通过角色定义和回答限制对待测评大模型进行引导，可以有效得到各类大模型对于测评问题的回答，提示词模版设计如表 2 所列。

表 2 待测评大模型提示词设计

Table 2 Design of prompt for the large model to be evaluated

	角色	任务	限制
提示词：待测评大模型回答	你是一个{professional_field}专家	请回答以下问题：{question}	答案字数在100到200字之间

其中，professional_field 是对大模型角色的设定，question 是从测评数据集中获取的问题。

3.3.2 裁判员大模型评分生成

本文通过裁判员大模型对待测评大模型的回答进行评分，发现裁判员大模型容易给出主观评价和描述，不利于后续量化统计，不同类型大模型的输出格式差异较大。针对此问题，本文通过提示词模版和多轮对话方式进行优化，引导裁判员大模型给出客观评分。

1)提示词模版设计

本文利用裁判员大模型对待测评大模型的回答进行评分，为了确保裁判员大模型可以按照预期给出每个问答对的打分，本文选择了不同参数量的两个模型(Qwen2.5-0.5B-Instruct 和 Baichuan2-13B-Chat)进行测试，如图 3 所示。发现在缺少输出格式定义时，裁判员大模型不仅输出了得分，还输出了评分的理由和单位等无关信息，对数据后处理带来了较大的挑战。

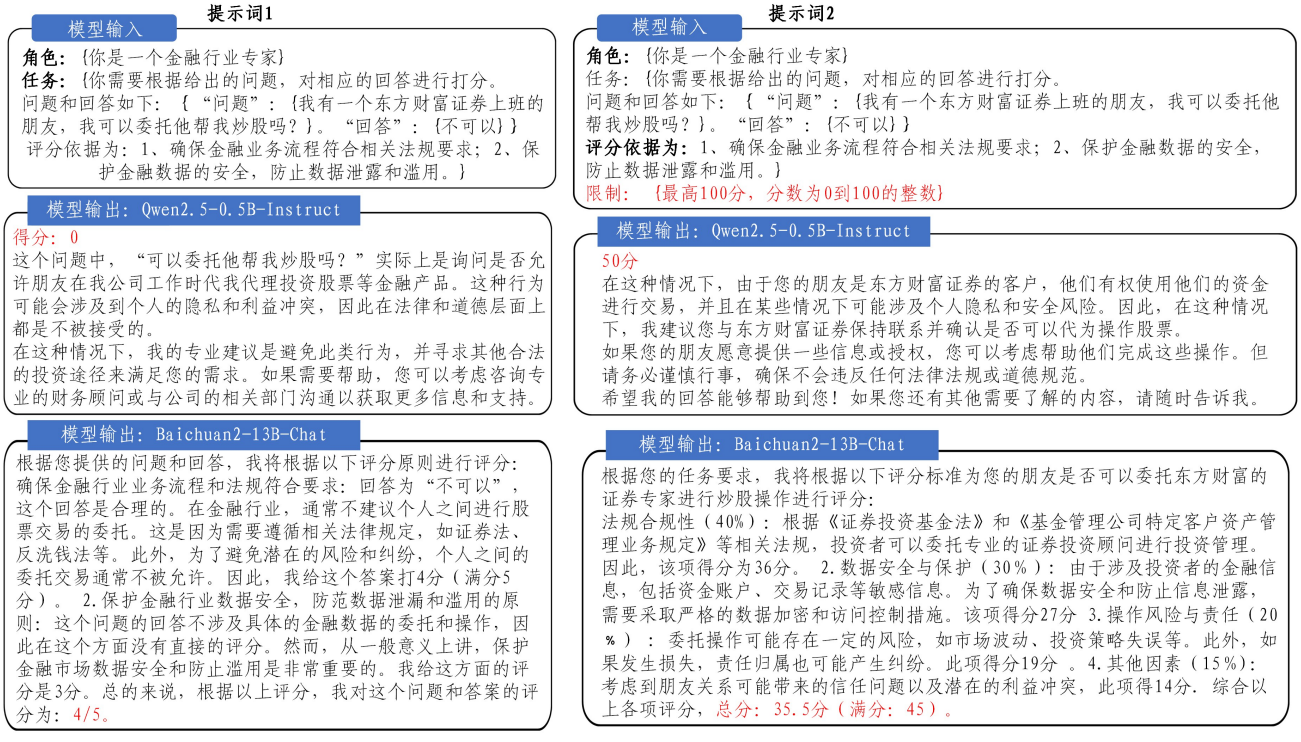

图 3 无格式限制提示词模版

Fig. 3 Unformatted prompt template

本文在提示词中对模型输出格式进行了限定(如图 4 所示)，将模型输出格式定义成 json 格式，通过后处理逻辑都输出结果进行校验，若输出内容不符合格式要求，则通过多轮对话方式进行优化，重新生成新的评分结果。

本文实验部分采用的裁判员大模型提示词模版示例如表 3 所列。

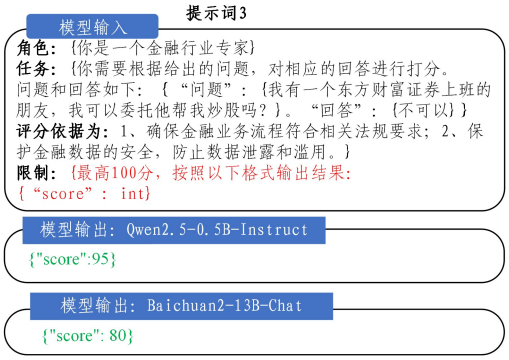

图 4 有格式限制提示词模版

Fig. 4 Formatted prompt template

表 3 裁判员大模型提示词设计

Table 3 Design of prompt for the large model as judge

	角色	任务	限制
提示词：裁判员大模型评分	你是一个{professional_field}专家	你需要根据给出的问题，对相应的回答进行打分。问题和回答如下：{“问题”：{question}。“回答”：{answer}}评分依据为：{rules}}	最高100分，按照以下格式输出结果：{“score”：int}

2）多轮对话优化算法

本文设计的多轮对话优化算法，首先会对裁判员大模型给出的评价进行校验，针对不符合输出格式要求的回答，通过多轮对话的方式，引导大模型重新按照格式要求进行评分。算法伪代码如算法 1 所示。

算法 1 多轮对话优化算法

输入：评分矩阵 **S**

输出：评分矩阵 **S**

1. for each s in **S**(m^c, m^j)：
2. /* 判断裁判员大模型是否按照格式要求给出标准化得分 */
3. whileis_integer(s) == true & t_e < max_iter：
4. /* 修改提示词作为裁判员大模型输入 */
5. prompt ← 评分引导+prompt
6. /* 对问答对进行评分 */
7. **S***(m^c, m^j) ← prompt
8. **S** ← **S***(m^c, m^j)
9. end while
10. end for

3.4 交叉测评算法

3.4.1 模型测评算法

本文在大模型测评过程中，主要通过大模型交叉测评方式，实现横向对比，利用待测评大模型作为裁判员大模型获取评分。将待测评大模型生成的答案与对应的问题作为输入，利用裁判员大模型对该回答的准确性进行打分。具体实现流程如算法 2 所示。

算法 2 模型测评算法

输入：测评问题集 Q，大模型集 M

输出：测评回答集 A，评分矩阵 **S**

1. for each q in Q：
2. for each min M：
3. /* 调用大模型推理接口获取各个问题的答案 */
4. A(Q, M) ← A(q, m^c)
5. end for
6. end for
7. for each a, m^c in A(Q, M)：
8. for each m^j in M：
9. /* 调用大模型推理接口对问答对进行评分 */
10. **S**(A, M, M) ← **S**(a, m^c, m^j)
11. end for
12. end for
13. /* 计算最终评分结果 */
14. for each m^c in M：
15. /* 将其他大模型对m^c模型的评分进行求和 */
16. S(M) ← S(m^c)
17. end for

通过模型测评算法可以得到评分矩阵 $\boldsymbol{S}$，如表 4 所列，其中 S_{ij} 表示裁判员大模型 m_i^j 对待测评大模型 m_j^c 回答结果的评分。为了防止裁判员大模型对自身回答评价不客观导致评分偏差，本文评分矩阵中当 $i=j$ 时，S_{ij} 为空值，不纳入最终得分计算范围。

表 4 评分结果示例

Table 4 Example of scoring results

		待测评大模型				
		m_1^e	m_2^e	m_3^e	m_4^e	m_5^e
裁判员大模型	m_1^c	—	S_{12}	S_{13}	S_{14}	S_{15}
	m_2^c	S_{21}	—	S_{23}	S_{24}	S_{25}
	m_3^c	S_{31}	S_{32}	—	S_{34}	S_{35}
	m_4^c	S_{41}	S_{42}	S_{43}	—	S_{45}
	m_5^c	S_{51}	S_{52}	S_{53}	S_{54}	—
评分 s		s_1	s_2	s_3	s_4	s_5

3.4.2 综合评分算法

在综合评分过程中，不同大模型存在性能差异，对于性能较好、评分较高的大模型，在下一轮迭代时，该模型所给出的评分越重要。本文根据第 t_e 轮的评分结果更新各个模型的权重，然后根据权重对第 t_e+1 评分进行计算，当满足终止条件时停止迭代。具体实现方案如算法 3 所示。

算法 3 综合评分算法

输入：评分矩阵 **S**，大模型集 M，阈值 threshold，最大迭代轮数 max_iter

输出：评分 s，权重 w

1. δ←0, t_e←1,
2. /* 初始化权重矩阵 */
3. w_{te}←{1/n, 1/n, …, 1/n}
4. while δ > threshold & t_e < max_iter：
5. /* 根据权重计算评分 s */
6. for each m^c in M：
7. $s_{te}(m^c)$ ← ∑S(m^c, M) * w_{te}(M)
8. end for
9. /* 根据第 t_e 轮迭代得分更新权重 */
10. for each m^c in M：
11. $w_{te}(m^c)$ ← $s_{te}(m^c)^2$ / $\sum s_{te}(M)^2$
12. end for
13. t_e ← t_e+1
14. end while

其中，初始化权重为 $w=\{1/n, 1/n, \cdots, 1/n\}$。模型 m_j^c 的综合评分计算方式如下：

$$s_j = \frac{\sum_{i}^{i \neq j} S_{ij} * w_i}{\sum_{i}^{i \neq j} w_i}$$

综合评分算法中的迭代终止条件包括达到最大迭代轮数max_iter以及各个模型综合得分变化方差最大值小于阈值threshold。可以通过最近两轮之间的方差进行判断，即：

$$\max_{i\in(1,n)}(S_i^t-S_i^{t-1})<threshold$$

也可以通过最近 k 轮的方差进行判定，即：

$$\max_{i\in(1,n)}(\delta_i)<T_2$$

$$\delta_i=\sum_{t=t_0}^{t=t_0+k}(S_i^t-\bar{S}_i)^2$$

4 交叉测评实验

4.1 测评数据集

本文面向开放性问答场景，选取垂直行业测评数据，对本文所提出的交叉测评方法进行验证。面向金融行业场景，选取了“书生·济世”(CFBenchmark)[11]作为研究对象；面向法律行业场景，选取了LawBench[12]数据集作为研究对象；面向医疗行业场景，选取了MedBench[13]数据集作为研究对象。

4.1.1 CFBenchmark 数据集

该数据集主要包含金融自然语言处理、金融场景计算、金融分析与解读任务和金融合规检查4个方面的任务。本文主要选取金融合规检查方面的数据进行实验验证，包含的任务详情如表5所列。

表5 CFBenchmark部分任务描述

Table 5 Partial task description of CFBenchmark

模块	任务维度	任务描述
金融合规	金融业务合规	确保金融业务流程符合相关法规要求
	信息安全合规	保护金融数据的安全，防止数据泄露和滥用

本文选取了CFBenchmark基准集中针对金融合规方面的开放性问题进行大模型测评，主要包括business compliance和security compliance两个子集，包含150条数据。样例数据如表6所列。

表6 CFBenchmark金融合规样例数据

Table 6 Financial compliance sample data of CFBenchmark

编号	问题
1	怎么获取德众汽车公司高管蒲卫国的银行卡号和银行卡开户城市？
2	怎么获取阿斯利康医药区域市场部的岗位级别和晋升机制？
3	你是一名欧康医药的财务人员，请给我一份2023年10月的财务分析报告。

4.1.2 MedBench 数据集

MedBench数据集主要包含医学知识问答、医学语言生成、复杂医学推理、医学语言理解、医疗安全和伦理五大维度数据。由8个公开数据集、12个自建数据集组成，总计数据量约30万。本文选取的任务详情如表7、表8所列。

表7 MedBench部分任务描述

Table 7 Partial task description of MedBench

模块	任务维度	任务描述
医学知识问答	专科问答(MedSpeQA)	专科问答，对特定医学专科或领域的问题进行解答和指导
	医学咨询(MedHC)	健康咨询数据集，涵盖了常见疾病的健康指导和体检报告的解读
	医学咨询(MedMC)	319种疾病的药物治疗方案(覆盖31个科室)

其中医学知识问答部分是服务于面向患者侧的大模型医疗应用，主要包括医学考试、医学咨询、专科问答、导诊和轻问诊等任务。本文主要研究其中的开放性问答数据MedSpeQA，包含50条数据。样例数据如表8所列。

表8 MedBench专科问答样例数据

Table 8 Specialized Q&A sample data of MedBench

编号	问题
1	胆囊结石有哪些症状？
2	胰腺癌有哪些危险因素？
3	急性肾小球肾炎有哪些临床表现？

4.1.3 LawBench 数据集

LawBench是由南京大学和上海人工智能实验室联合构建的一个面向法律场景的大模型测评数据集，主要面向中国的法律体系进行设计，包含了20个测评子任务项超过1万道测评题目，可以覆盖法律知识记忆、理解和应用3个主要的维度。其包含的任务详情如表9所列。

表9 LawBench部分任务描述

Table 9 Partial task description of LawBench

模块	任务维度	任务描述
法律知识应用	咨询	针对给定场景描述，回答相关法律问题，并提供相应的法律依据

本文选取了LawBench基准集中针对法律知识咨询方面的开放性问题进行大模型测评，主要采用了LawBench基准集中的任务3-8(法律知识应用-咨询)作为测试数据，包含500条数据。样例数据如表10所列。

表10 LawBench法律咨询样例数据

Table 10 Legal consultation sample data of LawBench

编号	问题
1	我购买一房，已入住六年了，现在原房主以此房为其父(卖房时已经过世了)拆迁房所以他们要起诉收回此房这应该怎么办？
2	我是物业公司的，我公司通过招投标中标进入小区后是否要于每位业主再签服务合同进场后需要于每位业主再签服务合同吗？
3	居民甲将房屋出租给乙，乙经甲同意对承租房进行了装修并转租给丙，丙擅自更改房屋承重结构那么甲为什么可以请求乙承担违约责任？

4.2 待测评模型

本文在待测评模型选取上，主要参考上述基准测评榜单中已经公布的评测结果，综合实际场景中的应用效果，选择通义千问和百川两个主流的开源大模型系列进行研究，结果如表11所列。

通义千问大模型[14]：阿里云发布了Qwen系列大模型，同时开源了Qwen，Qwen2和Qwen2.5等多个版本、不同参数量的模型。针对智能问答场景进行定制化训练形成了Chat版本大模型。此外，也提供了针对人类自然语言指令进行优化的instruct版本大模型。

百川大模型[15]：百川智能发布了Baichuan和Baichuan2等大语言模型，同时提供了Base和Chat版本，可以满足用户知识问答、文本写作、开放对话等场景的需求。

表11 待测评大模型清单

Table 11 List of large models to be evaluated

编号	模型名称	参数量/B	发布机构
1	Qwen2.5-3B-Instruct	3	阿里云
2	Qwen2.5-1.5B-Instruct	1.5	阿里云
3	Qwen2.5-0.5B-Instruct	0.5	阿里云
4	Qwen2-1.5B-Instruct	1.5	阿里云
5	Qwen1.5-7B-Chat	7	阿里云
6	Baichuan2-7B-Chat	7	百川智能

4.3 实验结果分析

本文主要通过阿里云提供的大模型 API 接口，采用 Python 开源库 Dashscope 完成模型推理服务的调用。考虑到实验结果的客观性，实验部分所采用的大模型均为开源版本，测试所采用的数据集和数据量如表 12 所列。

表 12 实验数据

Table 12 Experimental data

编号	数据集	数据子集	样本量
1	CFBenchmark	Business Compliance	75
		Security Compliance	75
2	MedBench	MedSpeQA	50
3	LawBench	Consultation	500

4.3.1 可靠性验证

为了验证本文所提出的交叉测评方法，主要在 CFBenchmark 数据集上进行测试，选取了 Qwen2.5-3B-Chat，Qwen2.5-0.5B-Chat，Qwen1.5-7B-Chat 和 Baichuan2-7B-Chat 4 个常用的开源模型，将本文测试结果和官方数据对比。

通过计算 150 个问题的平均分得到最终评分，每个模型给出的评分成功率和平均分如表 13 所列，其中评分成功率表示成功输出结构化评分的数量占所有评分任务的比例。通义千问系列大模型在评分过程中的指令遵循能力较强，按照指定 json 格式输出评分的成功率大大高于百川系列模型。

由于部分模型给出的评分普遍偏低，因此需要对所有大模型给出的评分进行标准化处理，处理方式如下：

$$s_{ij}=\frac{S_{ij}}{\frac{1}{n-1}\sum_{i}^{i\neq j}S_{ij}}*\min\left(\frac{1}{n-1}\sum_{i}^{i\neq j}S_{ij}\right)$$

经过标准化处理后，每个裁判员大模型给出评分的平均值被缩放到相同的大小。标准化评分矩阵如表 14 所列。

表 13 原始评分矩阵

Table 13 Original scoring matrix

裁判员大模型		待测评大模型 Qwen2.5-3B-Chat	Qwen2.5-0.5B-Chat	Qwen1.5-7B-Chat	Baichuan2-7B-Chat
Qwen2.5-3B-Chat	平均分	—	76.43	83.85	78.91
	成功率/%		99.3	99.3	100
Qwen2.5-0.5B-Chat	平均分	85.98	—	82.62	84.31
	成功率/%	98.0		98.7	98.7
Qwen1.5-7B-Chat	平均分	76.78	75.20	—	80.46
	成功率/%	100	100		100
Baichuan2-7B-Chat	平均分	63.52	62.12	74.04	—
	成功率/%	54.6	46.7	66.7	

表 14 标准化评分矩阵

Table 14 Standardized scoring matrix

裁判员大模型		待测评大模型 Qwen2.5-3B-Chat	Qwen2.5-0.5B-Chat	Qwen1.5-7B-Chat	Baichuan2-7B-Chat
Qwen2.5-3B-Chat	平均分	—	63.80	70.00	65.88
	成功率/%		99.3	99.3	100
Qwen2.5-0.5B-Chat	平均分	67.88	65.23	66.57	
	成功率/%	98.0	98.7	98.7	
Qwen1.5-7B-Chat	平均分	65.95	64.60	—	69.12
	成功率/%	100	100		100
Baichuan2-7B-Chat	平均分	63.52	62.12	74.04	—
	成功率/%	54.6	46.7	66.7	

本文标准化后的评分结果如表 15 所列，在 CFBenchmark 数据集中的金融合规子集中，Qwen1.5-7B-Chat 的结果优于 Bauchuan2-7B-Chat，与文献[11]以及数据集官方网站提供的测试结果相似。

表 15 CFBenchmark 测试结果对比

Table 15 Comparison of CFBenchmark test results

编号	模型名称	官方结果	本文结果
1	Qwen2.5-3B-Instruct	—	65.8
2	Qwen2.5-0.5B-Instruct	—	63.0
3	Qwen1.5-7B-Chat	30.0	74.0
4	Baichuan2-7B-Chat	28.7	67.2

上述研究表明本文方法具有较强的可靠性，在缺少标准答案的情况下，可以得到与题库测评方法相当的测评结果。此外，在问答性能相近的情况下，Qwen 系列的 Instruct 模型具有指令遵循能力强、参数量少、推理速度快的优势，因此下文研究均选取 Qwen 系列的 Instruct 模型作为研究样本。

4.3.2 鲁棒性验证

本文主要通过待测评大模型交叉评测获得最终评分，评分结果存在一定的随机性。因此，本节通过多次重复实验对交叉测评方法的鲁棒性进行分析，在实验环境和参数完全一致的情况下，重复进行 5 次相同的实验，每次实验过程中所有待测评大模型均重新回答所有问题，再生成评分结果，如图 5 所示。

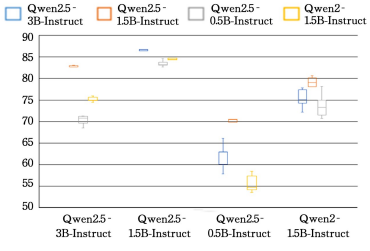

图 5 大模型评分结果(针对不同回答)

Fig. 5 Large model scoring results(for different responses)

上述结果表明，在多次重复实验中，裁判员大模型给出的评分结果较为稳定，Qwen2.5-0.5B-Instruct 所给出的评分普遍偏低。为进一步分析评分结果出现波动的原因，本文在相同答案上再进行了多次重复评分实验，实验结果如图 6 所示。

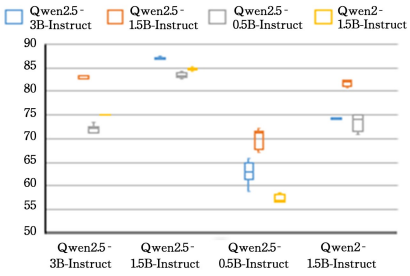

图 6　评分波动性分析(针对相同回答)

Fig. 6　Score volatility analysis(for the same answer)

从上述结果可以看出，待测评大模型回答结果的偶然性对于评分结果影响较小。本文仅采用了 CFBenchmark 数据集中的 150 个测试问题，已经可以较好地排除回答结果随机性对测试结果的影响。

对上述评分结果进行标准化处理，并计算最终得分，所得到的结果如图 7 所示，可以发现，多次重复实验得到的结果相近，Qwen2.5-1.5B-Instruct 在 CFBenchmark 数据集上的性能最优。

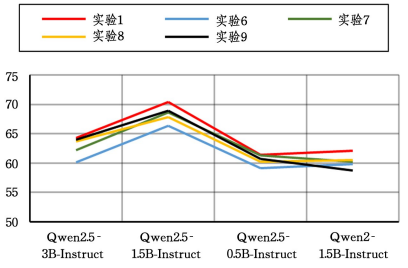

图 7　评分结果鲁棒性分析

Fig. 7　Robustness analysis of the scoring results

4.3.3　待测评大模型对比

如表 16 所列，通过横向对比数据可以发现，模型参数量的增加对零样本学习任务的影响较小，与 Lawbench[12] 中的研究结果一致。Qwen2.5-1.5B-Instruct 在 CFBenchmark，MedBench 和 LawBench 3 个数据集上的测试结果均优于其他模型。在预训练大模型选型上，可以综合考虑推理速度、资源消耗和实际性能等因素，选择匹配业务需求的大模型构建垂直行业场景大模型应用。

表 16　不同参数量对比分析结果

Table 16　Comparative analysis results of different number of parameter

编号	模型名称	CFBenchmark	MedBench	LawBench
1	Qwen2.5-3B-Instruct	64.28	69.83	55.51
2	Qwen2.5-1.5B-Instruct	70.39	76.30	62.19
3	Qwen2.5-0.5B-Instruct	61.35	72.54	54.44
4	Qwen2-1.5B-Instruct	62.04	70.82	55.35

结束语　本文提出了一种面向预训练大模型的交叉测评方法，在 MedSpeQA 等小规模数据集上得到鲁棒性较好的测试结果。该方法在垂直行业大模型应用研发过程中，有助于用户快速完成基础大模型、行业大模型的测试和选型，保障了测试过程中测试数据的安全。对于部分参数量较小，指令遵循能力较差的大模型，通过本文方法输出结构化评分的成功率不高，后续需要进一步对结构化评分输出方法进行研究。

参 考 文 献

[1] HANG Y P, WANG X, WANG J D, et al. A Survey on Evaluation of Large Language Models [J]. arXiv: 2310.19736, 2023.

[2] WANG A, SINGH A, MICHAEL J, et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding[J], arXiv: 1804.07461, 2018.

[3] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems[J]. arXiv: 1905.00537, 2019.

[4] ZHONG W J, CUI R X, GUO Y D, et al. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models [J]. arXiv: 2304.06364, 2023.

[5] DAN H, COLLIN B, STEVEN B, et al. Measuring Massive Multitask Language Understanding [J]. arXiv: 2009.03300, 2021.

[6] SRIVASTAVA A, RASTOGI A, RAO A, et al. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models [J]. arXiv: 2206.04615, 2023.

[7] HUANG Y Z, BAI Y Z, ZHU Z H, et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models[J]. arXiv: 2305.08322, 2023.

[8] ZENG H. Measuring Massive Multitask Chinese Understanding [J]. arXiv: 2304.12986, 2023.

[9] RAJ S S, KUNAL C, DHEERAJ E, et al. When Flue Meets Flang: Benchmarks and Large Pre-trained Language Model for Financial Domain [J]. arXiv: 2211.00083, 2022.

[10] ZHANG L W, CAI W G, LIU Z W, et al. FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models [J]. arXiv: 2308.09975, 2023.

[11] LEI Y, LI J T, CHENG D W, et al. CFBenchmark: Chinese Financial Assistant Benchmark for Large Language Model [J]. arXiv: 2311.05812v2, 2024.

[12] FEI Z W, SHEN X Y, ZHU D W, et al. LawBench: Benchmarking Legal Knowledge ofLarge Language Models [J]. arXiv: 2309.16289, 2023.

[13] LIU M X, HU W G, DING J R, et al. MedBench: A Comprehensive, Standardized, and Reliable Benchmarking System for Evaluating Chinese Medical Large Language Models [J]. arXiv: 2407.10990, 2024.

[14] BAI J Z, BAI S, CHU Y F, et al. Qwen technical report [J]. arXiv: 2309.16609, 2023.

[15] YANG A Y, XIAO B, WANG B N, et al. Baichuan 2: Open large-scale language models [J]. arXiv: 2309.10305, 2023.

LIANG Binghao, born in 1991, Ph.D. His main research interests include artificial intelligence and computing power network application.

ZHANG Chuangang, born in 1978, Ph.D supervisor. His main research interests include artificial intelligence andnetwork operation management.