计算机科学
COMPUTER SCIENCE

基于选择题的语言模型学科知识评估方法

熊卓帜, 顾洲洪, 冯红伟, 肖仰华

引用本文

熊卓帜, 顾洲洪, 冯红伟, 肖仰华. 基于选择题的语言模型学科知识评估方法[J]. 计算机科学, 2025, 52(10): 201-207.

XIONG Zhuozhi, GU Zhouhong, FENG Hongwei, XIAO Yanghua. Subject Knowledge Evaluation Method for Language Models Based on Multiple ChoiceQuestions [J]. Computer Science, 2025, 52(10): 201-207.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240800148

基于选择题的语言模型学科知识评估方法

熊卓帜　顾洲洪　冯红伟　肖仰华
复旦大学计算机科学技术学院　上海 200433
(xiongzz21@m.fudan.edu.cn)

摘　要　基于选择题(Multiple Choice Question,MCQ)形式的预训练语言模型学科知识评估方法虽然能够快速、量化地评估模型的知识,但其可靠性受到选项顺序和选项长度等无关因素的干扰,存在稳健性问题。为解决这一挑战,首先提出了一个基于MCQ的预训练语言模型学科知识评测分析框架。该框架将MCQ评测方法形式化为提示和解析两个模块,系统探索了不同类别的MCQ模型评测方法对评测结果的影响。在中英文学科知识评测数据集上的实验分析了不同提示和解析方法的稳健性。基于分析结果,提出了一种改写增强的解析方法,该方法通过引入预训练语言模型对模型回复进行改写,有效解决了传统基于规则解析的方法在处理非标准回复时的局限性。通过结合改写和规则解析,不仅提高了答案提取的准确性,还增强了评测过程的稳健性,为语言模型评测提供了一种新的有效途径。

关键词:语言模型;学科知识;选择题评估
中图分类号　TP391

Subject Knowledge Evaluation Method for Language Models Based on Multiple Choice Questions

XIONG Zhuozhi,GU Zhouhong,FENG Hongwei and XIAO Yanghua
School of Computer Science,Fudan University,Shanghai 200433,China

Abstract　Subject knowledge evaluation methods for pre-trained language models based on Multiple Choice Questions(MCQ) offer rapid,quantitative evaluation of model knowledge. However,their reliability is compromised by irrelevant factors such as option order and length,raising robustness concerns. To address this challenge,an analytical framework for evaluating subject knowledge of pre-trained language models using MCQ is proposed. This framework formalizes MCQ evaluation into two modules:prompting and parsing,systematically investigating the impact of various MCQ evaluation methods on evaluation outcomes. The robustness of different prompting and parsing techniques is analyzed through experiments on Chinese and English subject knowledge evaluation datasets. Based on these findings,a rewriting-enhanced parsing method is introduced that employs pre-trained language models to rewrite model responses,effectively overcoming the limitations of traditional rule-based parsing when handling non-standard replies. By integrating rewriting and rule-based parsing,this approach enhances both answer extraction accuracy and evaluation process robustness,offering a novel and effective strategy for language model evaluation.

Keywords　Language model,Subject knowledge,Multiple choice evaluation

1　引言

为有效且准确地评估预训练语言模型的性能,使用选择题(MCQ)来提高模型评测的可控性已成为一种普遍趋势[1-3]。选择题评测将模型的输出限制在预定义的选项集合中,减少了对参考文本的依赖,从而提高了评测的可控性和准确性。

在选择题评测中,一个关键挑战是预训练语言模型对提示模块和解析模块表现出的敏感性。即使是对提示和解析方法的微小调整,也可能对模型性能产生显著影响,进而导致评测结果的变化[4-5]。这种对提示和解析方法的敏感性引发了两个关键问题:不同提示和解析方法的差异及其各自的优点是什么?如何才能增强评测的稳健性?

鉴于预训练语言模型的实际应用场景,本研究提出提示和解析方法应满足3项关键标准。1)应答能力:定义为模型在使用此种提示-解析策略时,回答出合理选择的能力,即根据输入的提示能够明确所选答案的能力。2)准确性:定义为模型在使用此种提示-解析策略时,回答出正确答案的能力。3)抗干扰性:抗干扰性定义为模型在使用此种提示-解析策略时,针对输入的微小扰动能够产生一致的回答的能力。

本研究将选择题(MCQ)的评测过程划分为两个关键模块:提示(Prompting)模块和解析(Parsing)模块(见图1)。提示模块负责将原始问题转换为适合语言模型处理的输入格式。解析模块则旨在从模型生成的输出中提取和解析出预测答案。本研究综合了6种主流提示方法和4种解析方法,对

到稿日期:2024-08-28　返修日期:2024-10-31
通信作者:肖仰华(shawyh@fudan.edu.cn)

多个当前主流的中英文语言模型进行了系统性实验。通过实验分析，本研究识别出了目前 MCQ 评测方法的主要瓶颈，并提出了一种基于预训练语言模型的改写方法，以提升 MCQ 评测的稳健性。

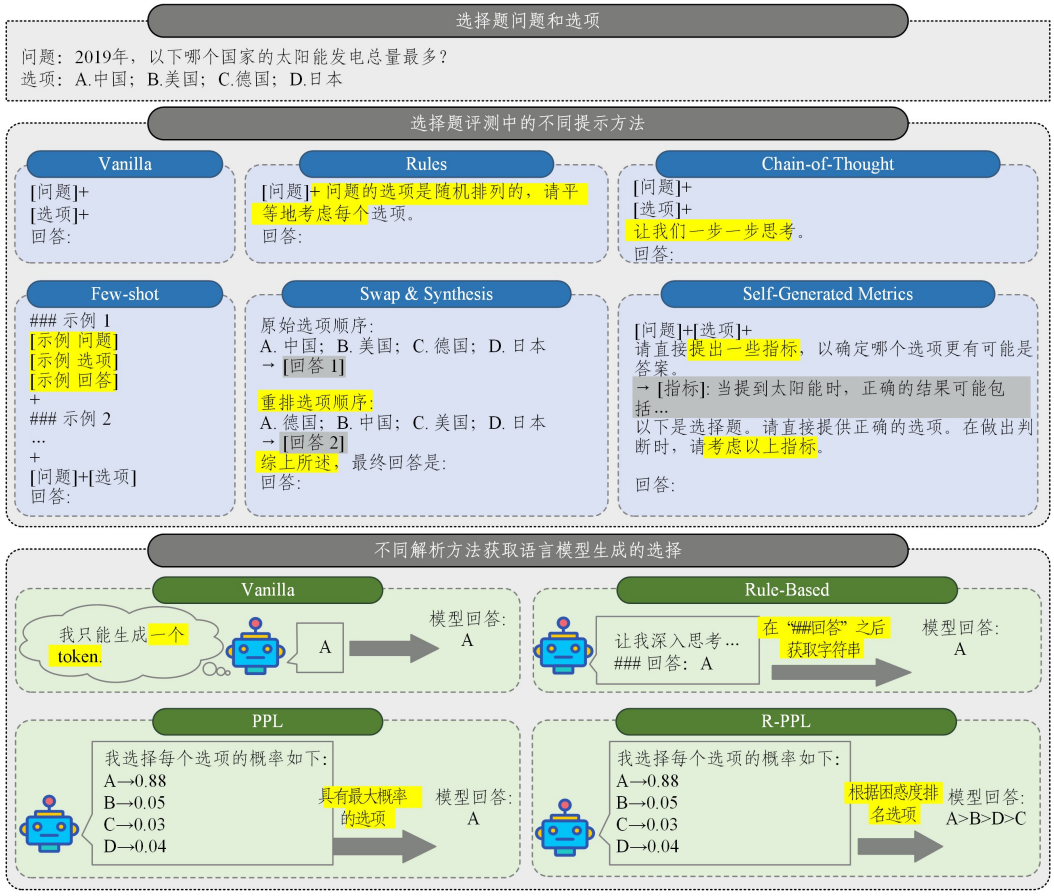

图 1 不同提示和解析方法的示例

Fig. 1 Examples of different prompting and parsing methods

本研究的主要贡献如下：

1）即使使用相同的模型和数据集，提示和解析方法的微小变化也可能导致显著不同的评测结果和模型排序。某些模型倾向于适应特定的提示-解析范式，而在更改提示-解析方法后，其性能有较大的差别。这表明部分模型在不同评测方法中缺乏稳健性，导致不同提示-解析方法的评测结果之间存在显著差异。

2）提示-解析方法之间的不对齐现象可能是导致评测结果不稳定的主要原因，同时这种现象也限制了评测性能的上限。

3）根据本文实验得出的结论，提出了一种改写模型初始输出的方法。该方法旨在弥补提示方法和解析方法之间的差异，从而提高模型的应答能力，并最终增强基于学科知识的预训练语言模型评测方法的稳健性。

2 相关工作

2.1 选择题形式的学科知识评估

近年来，很多研究工作致力于为语言模型的知识评测提供 MCQ 形式的数据集，这些数据集分别关注不同维度的知识。GAOKAO-bench[6] 和 ARC[7] 分别基于中国高考题目和小学阅读理解题目衡量语言模型的语言理解能力和逻辑推理能力，并将其与人类的表现进行对比。MMLU[1]，C-eval[2] 和 CMMLU[3] 也采用不同语种的选择题形式评估语言模型的世界知识。

2.2 基于选择题的评测方法的稳健性

尽管 MCQ 形式的评测方法有着诸多优点，但许多工作[4-5,8-9]也发现，MCQ 形式的评测方法存在不稳健的问题。Zheng 等[5]发现模型在面对问题相同但选项顺序不同的选择题时表现出不一致性，并提出一种综合多个选项顺序的推理结果以增强评估的稳健性的方法。但这种方法需要被评测模型针对同一道题进行多次推理，增加了评测模型的推理成本。Scherrer 等[8]研究大规模语言模型的道德信念，发现在道德上有争议的场景中某些模型的选择表现出较大的不确定性，其观点在不同题目中存在较大差异。但此工作仅仅使用规则将模型输出与选项对应，并没有探究以字符概率为基础的解析方法。Hu 等[9]比较了直接使用字符概率和元语言的提示（Metalinguistic Prompting）在评估大规模语言模型的语言知识方面的差异。研究发现，大规模语言模型通过提示方法得出的评测答案的准确性通常低于直接从模型的字符概率分布中得出结果的方法。此外，元语言的提示方法和基于字符概率的方法得出的答案在分布上具有不一致性。所以，Hu 提出尽量不要直接使用生成文本的方法代替基于字符概率的评测方法，因为其很难将以自由形式生成的文本与所选选项相对应，从而无法稳健地评估模型的性能。

3 评估方法

3.1 评价指标

本节提出一个使用选择题评测预训练语言模型性能的理论框架。该框架主要基于上文提出的评价标准和提示-解析范式。评测框架的总体目标是最大化预训练语言模型的应答能力、准确性和抗干扰性。给定一个问题 q 和一组可能的答案 $A=\{a_1,a_2,\cdots,a_n\}$,定义 pro 是选定的提示词,a' 是语言模型预测的答案,a^* 是正确答案。根据以上定义,将这些标准形式化地定义为:

1)应答能力。给定提示词 pro、解析方法 par、问题 q 和候选答案集 A,最大化模型生成合理候选答案在 A 中的概率 P_r:

$$\text{argmax}\ P_r(a\in A\mid q,A,pro,par) \tag{1}$$

2)准确性。最大化解析后的答案是正确答案的概率 P_a:

$$\text{argmax}\ P_a(a'=a^*\mid q,A,pro,par) \tag{2}$$

3)抗干扰性。最小化同一问题的不同问法的结果之间的差异。其中,ϵ 表示微小扰动,P_{valid0} 表示原问法可得到有效回答的概率,P_{valid1} 表示微小扰动后问法可得到有效回答的概率,P_{true0} 表示原问法可得到正确回答的概率,P_{true1} 表示微小扰动后问法可得到正确回答的概率。

$$\begin{aligned}\text{argmin}\ [&|P_{\text{valid0}}(a'\in A\mid pro,q,A)-P_{\text{valid1}}(a'\in A\mid pro+\epsilon,q,A)|+\\&|P_{\text{true0}}(a'=a^*\mid pro,q,A)-P_{\text{true1}}(a'=a^*\mid pro+\epsilon,q,A)|]\end{aligned} \tag{3}$$

3.2 提示模块

提示模块旨在引导预训练语言模型生成最准确和相关的回答。提示模块的目标是构建一个提示 $pro(q,A)$,以最大化预训练语言模型从集合 A 中生成正确答案的概率,可形式化地表示为:

$$\begin{aligned}pro(q,A)=\arg\max_{pro}[&P(a'\in A\mid pro,q,A)+\\&P(a'=a^*\mid pro,q,A)]\end{aligned} \tag{4}$$

其中,P 表示所有可选提示的集合;C 表示预训练语言模型生成的内容;而选定的 $pro(q,A)$ 是由前文所述标准决定的;a' 代表预训练语言模型所做的选择,其需要与解析模块相结合,以确定预训练语言模型生成的最终答案;a^* 是正确答案。

3.3 解析模块

解析模块的目标是从预训练语言模型生成的内容中解析并提取出其所做出的选择。这个过程可以形式化为:

给定生成的内容 C,解析模块的目标是提取预训练语言模型输出的答案 a',最大化其成为集合 $A=\{a_1,a_2,\cdots,a_n\}$ 中可能的候选答案之一的概率。然而,如果 C 中没有 A 中的答案,则不应错误地将任何其他内容解析为答案。形式化地表示为:

$$\begin{aligned}a=\arg\max_{par}[&I(a'\in A)\cdot P(a'\mid C,par)+I(a'\notin A)\cdot\\&P(\varnothing\mid C,par)]\end{aligned} \tag{5}$$

其中,$P(a'\in A\mid C,par)$ 是给定生成内容 C 的情况下,答案 a' 在候选答案集 A 中的概率;$I(\cdot)$ 是指示函数,当括号内条件为真时值为 1,否则为 0;$\varnothing$ 为空集,代表 C 无法被解析时,不应错误地将任何其他内容解析为答案。

3.4 基于改写增强的解析方法

在选择题的语言模型评测领域,目前主流评测榜单普遍采用基于规则的解析方法(Rule-based Parsing)。该方法需要人类专家制定一系列规则并将其纳入规则库。在解析过程中,评测框架利用规则库中的规则对模型生成的文本进行正则表达式匹配,从而提取出模型所选择的答案。

基于规则的解析(Rule-based Parsing)方法依赖预定义规则来解析和理解模型输出。这种方法在处理特定类型的提示时表现卓越,因为它能够直接映射到规则集,从而快速准确地提取所需信息。然而,这种方法的局限性在于其对规则的高度依赖。当模型的回复与预设规则不匹配时,即出现不对齐现象,该方法的有效性将大幅降低。

为了克服这一局限,本研究提出了一种改进方案:利用预训练语言模型对基于规则的解析方法进行优化。这种方法有效解决了传统基于规则的方法在从模型回复中提取答案时的问题,提高了信息提取的灵活性和准确性。

该方法的实施步骤如下。1)预处理:将问题和模型的初始回复输入至改写模型中,两者以换行符分隔。2)改写:改写模型对初始回复进行优化,以增强其与模型答案的相关性。本方法在输入中加入显式指令,要求改写模型明确地将输出归类为 4 个预设选项之一或标记为"无法确定"。3)基于规则的解析:经改写的回复随后传输至基于规则的解析模块。由于回复质量得到提升,答案更加明确,该解析方法能更准确地识别回复中的关键信息并提取答案。4)结果输出:最终输出经过改写和基于规则解析处理后的答案。这种方法不受具体规则限制,使得评测结果更具稳健性。

4 实验分析

4.1 实验设置

本工作主要在基于选择题的中英文知识评测数据集 MMLU[1],C-Eval[2] 和 CMMLU[3] 上进行实验。实验主要使用 LLaMA2 系列[10]、ChatGLM 系列[11-12]、GPT-3.5[13] 和 GPT-4[14],以确保对不同模型架构进行全面评测。除非特别说明,所有生成参数都采用贪婪解码,temperature 设为 0。

4.2 基线方法

1)提示方法

(1)Vanilla 方法在提示词中要求预训练语言模型直接从可能的答案集合中生成所选选项的字符,而不回答其他内容。

(2)思维链(Chain-of-Thoughts,CoT)提示方法[15] 通过要求预训练语言模型在生成最终答案之前先生成推理过程来提升模型在任务中的性能。

(3)少样本演示(Few-Shot Demonstration,FSD)是小样本学习的一种形式,它通过在问题前面添加若干示例进行演示来辅助模型回答问题。

(4)Rules 提示方法[16] 在提示中明确列出一些需要预训练语言模型遵循的规则。

(5)Swap&Synthesis(Swap)提示方法[16-17] 通过交换提示中选项的顺序来缓解提示模块中的位置偏见问题。

(6)Self-Generated Metrics(SGM)提示方法[16] 通过让模型自己生成一组指标来指导预训练语言模型的输出生成。

2)解析方法

(1)Vanilla 解析方法使用默认的解码策略,让预训练

语言模型仅输出一个字符，将其作为预训练语言模型的输出答案 a'。

(2)Rule-based Parsing(RP，基于规则的解析方法)[2-3]通过使用正则表达式匹配从生成的文本中提取模型的预测结果。

(3)PPL 解析方法[18]将每个选项分别附加在问题后，选择模型生成最低困惑度的选项a'。

(4) R-PPL 解析方法[18]是 PPL 方法的扩展，其定义为正确选项在 4 个选项中的 MRR(Mean Reciprocal Rank)。

4.3 实验结果

实验测试了不同提示和解析方法组合的准确性得分，结果如表 1 所列，其中 R-PPL 分数为 MRR，其他解析方法的分数为 Accuracy。表中数据为在 0-shot 设置下，提示-解析方法实验的结果。对于应答能力的衡量，实验在不同提示和解析设置下计算有效回答率(Valid Response Rate，VRR)。有效回答率定义为从模型输出中可被对应解析方法正常解析为候选答案集中答案的比例。图 2 展示了在 MMLU，CMMLU，C-Eval 上，采用同样的提示-解析方法，不同模型间的准确率的 spearman 相关性。图 3 展示了在 MMLU，CMMLU，C-Eval 上，采用同样模型，不同提示-解析方法间的准确率的 spearman 相关性。图 4 展示了在不同提示和解析设置下各种模型的有效回答率。图中 SGM 代表 Self-Generated Metric。图例的格式为提示方法|解析方法。图中的数值为各个模型在 MMLU，CMMLU 和 C-Eval 上性能的平均 VRR。图 5 展示了不同解析方法下，GPT-3.5-Turbo 在 MMLU，CMMLU 和 C-Eval 数据集上针对思维链提示方法进行改写后的有效回答率。

表 1 提示-解析方法实验结果

Table 1 Result of promting-parsing experiment

Dataset	CMMLU				MMLU				C-eval			
Prompting	Vanilla											
Parsing	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL
ChatGLM-6B	40.0	40.8	39.8	0.63	40.1	40.4	40.8	0.63	37.1	40.6	41.5	0.64
ChatGLM2-6B	50.8	49.6	49.3	0.69	46.4	46.4	47.4	0.68	53.3	53.8	53.6	0.72
ChatGLM3-6B	52.7	55.5	54.8	0.73	48.1	49.8	51.5	0.70	68.1	59.9	67.9	0.81
ChatGLM3-6B-base	67.4	69.4	69.9	0.82	62.1	62.0	62.4	0.77	67.9	59.9	67.9	0.81
LLaMA2-7B	28.9	29.5	30.0	0.56	35.3	29.6	36.0	0.60	30.3	34.8	35.4	0.60
LLaMA2-7B-Chat	18.9	30.7	32.0	0.57	46.1	26.1	46.8	0.67	30.6	34.8	35.9	0.60
LLaMA2-13B	33.9	36.3	37.3	0.61	48.6	26.8	48.7	0.68	37.4	37.0	37.4	0.61
LLaMA2-13B-Chat	10.7	32.6	36.0	0.59	46.4	25.7	51.0	0.70	22.7	32.1	32.5	0.57
GPT-3.5-turbo	53.0	50.4	—	—	68.2	68.3	—	—	50.2	49.5	—	—
GPT-4-turbo	71.6	72.3	—	—	81.3	81.3	—	—	68.0	68.9	—	—
Prompting	Rules											
Parsing	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL
ChatGLM-6B	38.6	33.1	34.9	0.62	38.8	40.3	42.2	0.64	38.5	41.4	42.1	0.64
ChatGLM2-6B	49.6	49.8	50.1	0.70	47.1	47.3	46.4	0.67	54.5	55.1	53.6	0.73
ChatGLM3-6B	65.9	66.2	66.5	0.80	52.3	51.5	51.0	0.69	67.6	59.7	66.9	0.80
ChatGLM3-6B-base	66.6	67.8	68.7	0.81	61.5	61.7	62.4	0.77	53.0	53.0	52.5	0.69
LLaMA2-7B	27.1	30.4	31.1	0.56	35.1	27.1	38.9	0.62	29.0	34.3	37.1	0.60
LLaMA2-7B-Chat	16.6	29.6	31.7	0.57	45.4	25.9	46.3	0.67	29.3	34.8	35.5	0.60
LLaMA2-13B	35.4	36.0	34.7	0.59	47.9	23.1	48.1	0.68	35.1	34.8	35.3	0.58
LLaMA2-13B-Chat	10.0	30.1	34.4	0.59	46.3	26.0	50.5	0.70	24.5	32.5	32.6	0.57
GPT-3.5-turbo	52.0	49.3	—	—	67.1	67.3			53.4	50.1	—	—
GPT-4-turbo	70.6	72.3	—	—	79.5	81.3		—	64.3	67.6	—	—
Prompting	Swap & Synthesis											
Parsing	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL
ChatGLM-6B	28.5	25.2	33.3	0.59	30.4	30.8	32.7	0.59	34.2	36.3	36.4	0.60
ChatGLM2-6B	31	33.1	34.9	0.6	30.8	47.3	42.2	0.64	33.5	35.5	37.4	0.62
ChatGLM3-6B	46.9	49.7	49.3	0.69	48.0	51.5	48.2	0.68	66.9	25.5	67.1	0.80
ChatGLM3-6B-base	67.6	67.7	67.8	0.80	61.1	62.1	62.1	0.76	65.4	65.6	65.6	0.79
LLaMA2-7B	26.7	27.1	28.5	0.55	28.6	26.8	30.8	0.52	33.3	34.3	34.9	0.58
LLaMA2-7B-Chat	29.1	30.5	30.6	0.56	44.1	25.0	45.2	0.65	33.1	34.5	34.7	0.58
LLaMA2-13B	33.9	36.3	37.3	0.61	38.6	26.3	42.1	0.64	31.3	31.4	34.5	0.57
LLaMA2-13B-Chat	23.5	30.7	29.9	0.56	40.4	24.6	42.6	0.65	26.7	31.6	32.8	0.57
GPT-3.5-turbo	53.0	50.4	—	—	62.4	65.2	—	—	44.2	45.5	—	—
GPT-4-turbo	69.7	69.5	—	—	79.1	80.5	—	—	67.9	69.1	—	—
Prompting	Self-Generated Metrics											
Parsing	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL	Vanilla	RP	PPL	R-PPL
ChatGLM-6B	31.8	27.9	38.4	0.62	30.4	30.8	32.7	0.59	41.5	41.7	41.9	0.63
ChatGLM2-6B	45.5	45.8	45.8	0.62	30.8	47.3	42.2	0.64	33.5	35.5	37.4	0.62
ChatGLM3-6B	51.8	51.8	52.2	0.71	48.0	51.5	48.2	0.68	66.9	25.5	67.1	0.80
ChatGLM3-6B-base	63.0	63.0	64.4	0.78	60.6	60.7	61.1	0.77	58.9	59.0	60.5	0.76
LLaMA2-7B	11.1	27.4	30.6	0.56	28.6	26.8	30.8	0.52	33.3	34.3	34.9	0.58
LLaMA2-7B-Chat	21.2	30.6	29.7	0.56	28.4	25.2	40.8	0.63	26.6	29.6	30.4	0.56
LLaMA2-13B	20.9	29.2	33.5	0.58	35.1	33.3	42.1	0.64	24.0	31.4	31.5	0.56
LLaMA2-13B-Chat	9.6	26.8	33.0	0.58	37.9	26.1	46.9	0.67	16.8	31.3	31.6	0.56
GPT-3.5-turbo	50.5	49.3	—	—	62.4	65.2	—	—	44.2	45.5	—	—
GPT-4-turbo	67.4	70.4	—	—	80.7	80.7	—	—	63.0	66.1	—	—

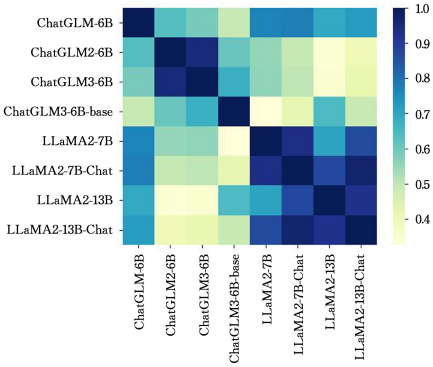

图 2 不同模型间的准确率的 spearman 相关性

Fig. 2 Spearman correlation of accuracy between different models

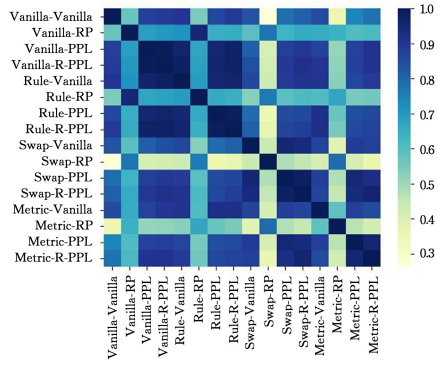

图 3 不同提示-解析方法间的准确率的 spearman 相关性

Fig. 3 Spearman correlation of accuracy between different prompting-parsing method

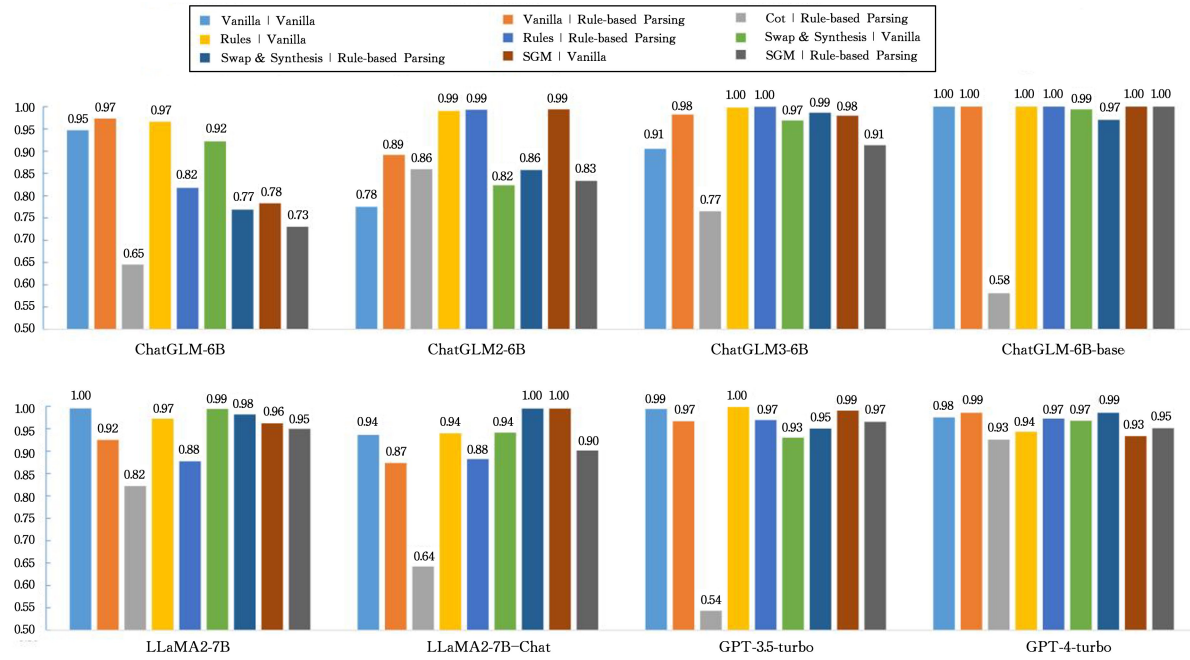

图 4 使用 Vanilla 和 Rule-based Parsing 方法的 VRR

Fig. 4 VRR using Vanilla and Rule-based Parsing methods

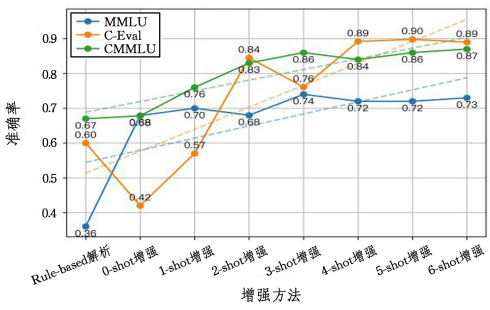

图 5 不同解析方法下,GPT-3.5-Turbo 针对思维链提示方法进行改写的 VRR

Fig. 5 VRR of GPT-3.5 Turbo in rewriting the CoT response under different parsing methods

4.4 实验分析

4.4.1 解析方法分析

1)基于规则的解析方法

基于规则的解析(Rule-based Parsing)并非普遍适用,而是仅适合特定的提示方法。如表 1 所列,该方法更适合于不依赖模型自行生成上下文的提示技术。在 Self-Generated Metrics 和思维链(CoT)提示方法中,基于规则的解析方法的表现逊色于 Vanilla 方法,但其与其他提示方法结合时却能展现出优越性。

基于规则的解析方法对模型性能有显著影响。如图 3 所示,该方法与其他解析方法的性能相关性较低,尤其是在使用 Swap & Synthesis 提示方法等需要模型输出长文本的提示方法时。表 1 进一步表明,对于不同的提示方法,基于规则的解析方法可在原始解析方法的基准上显著提高或降低性能,这主要取决于具体规则与相应模型和提示方法的集成程度。因此,Rule-base Parsing 在模型评测中相当于一把“双刃剑”,需要谨慎选择用于解析的规则。

2)基于 PPL 的解析方法

基于 PPL 解析方法的性能相对稳定。如表 1 所列,当使用模型自身生成的内容作为输入时,PPL 和 R-PPL 表现出一致的稳定性。这种稳定性可能源于以下原因:当提示方法的

形式与预训练语料库的分布存在显著差异时，模型倾向于将正确答案的字符排列在较后的位置。然而，PPL 和 R-PPL 仅考虑 4 个候选选项的字符之间的相对位置，从而有效缓解了这一问题。

3）模型间对解析方法的偏好一致性

提示/解析方法的性能在同一模型系列中具有相似性。实验通过固定提示方法和解析方法，来计算模型之间性能表现的 spearman 相关性。如图 2 所示，同一模型系列之间性能的相关性更为显著，例如 LLaMA2 系列和 ChatGLM3 系列。值得注意的是，未经指令微调的 Base 版本模型与经过指令微调的 Chat 版本模型在预测结果上表现出较高的相似性。具体而言，在 LLaMA2 中，Base 和 Chat 模型之间的差异很小；而在 ChatGLM3 系列中，尽管 Base 和 Chat 模型之间的性能差异略大，但这种差异仍然小于不同基座模型之间的差异。

4.4.2 提示方法分析

1）CoT 提示方法

CoT 提示方法与 Rule-based Parsing 方法组合可能会使有效回答率下降。如图 4 所示，CoT 提示方法通常会降低模型生成可解析内容的概率。这一发现表明，尽管 CoT 通过引入推理过程可以提升模型的整体性能，但由于生成的内容难以被规则准确解析，在选择题评测中可能无法充分体现出性能的提升。因此，在使用选择题进行模型评测时，需要谨慎考虑并精心设计解析方法，以确保准确捕捉模型的真实性能。

2）简单提示方法

参数量少的模型更加适用于简单的提示方法。如表 1 所列，Vanilla 提示和 Rules 提示的性能优于 SGM 和 Swap & Synthesis 提示方法。这些方法的共同特点是：都会要求模型生成更多的中间推理步骤。这表明，Vanilla 和 Rules 提示方法比 SGM 和 Swap & Synthesis 提示方法更简单，它们可能更加适用于小模型。

3）Rules 提示方法

Rules 提示方法是相对稳定的提示方法。如表 1 所列，Rules 提示方法是最稳定的，无论是在不同的模型中还是在不同的解析方法中，都保持着较为一致的性能排名，展现出最高的稳定性。相比之下，CoT 和 SGM 方法的表现可能会受到不同的解析方法的影响。

4.4.3 改写增强的解析方法分析

1）零样本改写增强的效果

零样本改写增强能提高有效回答率，但其效果与回答的语言类型密切相关。与基于规则的解析方法相比，零样本增强方法在 3 个数据集的其中两个上，有效回答率得到了提升，其中最显著的提升出现在英文知识评测数据集 MMLU 中。在 MMLU 数据集上，零样本增强方法实现了最大幅度的有效回答率提升，这表明即使在没有额外训练样本的情况下，改写模型仍能生成高质量的英文改写。然而，在其他数据集上，零样本增强的效果与基于规则的解析方法相当或略低，这反映了不同数据集的复杂性以及模型泛化能力的局限性。基于 GPT-3.5-turbo 和 GPT-4 的零样本改写增强方法主要在英文数据集中提高了评测方法的稳健性。

2）少样本上下文学习的影响

增加少量示例的上下文学习能显著提升改写模型在答案改写和总结方面的能力。研究结果表明，随着示例数量从 1 个增加到 6 个，模型在所有数据集上的有效回答率呈现普遍上升趋势。这一现象说明，即使是少量的示例也能帮助改写模型更好地理解和适应答案提取与改写任务，从而提高回答的质量和稳健性。这种趋势在 MMLU 数据集上尤为明显，进一步证实了增加示例数量与有效回答率提升之间的积极相关关系。

结束语 本研究旨在分析并提高选择题形式下预训练语言模型学科知识评估的稳健性，主要从应答能力、准确性和抗干扰性 3 个维度进行考察。本研究基于多语言数据集对多语言模型进行了实验，发现在基于选择题的预训练语言模型评测框架中普遍存在提示-解析方法不对齐的问题。针对这一现象，提出了一种利用预训练语言模型进行输出改写的方法，以使原有模型的输出更好地适配解析方法。实验结果表明，该方法能够有效缓解提示方法与解析方法之间的不对齐现象，提高评估框架对模型输出答案的解析能力，从而更准确地评测预训练语言模型中的学科知识。

参 考 文 献

[1] HENDRYCKS D, BURNS C, BASART S, et al. Measuring Massive Multitask Language Understanding [C] // International Conference on Learning Representations. 2021.

[2] HUANG Y, BAI Y, ZHU Z, et al. C-eval: A multi-level multi-disciplinechinese evaluation suite for foundation models [C] // Advances in Neural Information Processing Systems. 2023.

[3] 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: 11260-11285.

[4] PEZESHKPOUR P, HRUSCHKA E. Large language models sensitivity to the order of options in multiple-choice questions [C] // Findings of the Association for Computational Linguistics: NAACL 2024. ACL, 2024: 2006-2017.

[5] ZHENG C, ZHOU H, MENG F, et al. Large language models are not robust multiple choice selectors [C] // The Twelfth International Conference on Learning Representations. 2024.

[6] ZHANG X, LI C, ZONG Y, et al. Evaluating the performance of large language models on gaokao benchmark [J]. arXiv: 2305.12474, 2023.

[7] CLARK P, COWHEY I, ETZIONI O, et al. Think you have solved question answering? try arc, the ai2 reasoning challenge [J]. arXiv: 1803.05457, 2018.

[8] SCHERRER N, SHI C, FEDER A, et al. Evaluating the moral beliefs encoded inllms [C] // Advances in Neural Information Processing Systems. 2024.

[9] HU J, LEVY R. Prompt-based methods may underestimate large language models' linguistic generalizations [J]. arXiv: 2305.13264, 2023.

[10] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models [J]. arXiv: 2307.09288, 2023.

[11] DU Z，QIAN Y，LIU X，et al. Glm：General language model pretraining with autoregressive blank infilling[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics ACL，2022：320-335.
[12] ZENG A，LIU X，DU Z，et al. Glm-130b：An open bilingual pretrained model [J]. arXiv：2210. 02414，2022.
[13] CHATGPT I. Introducingchatgpt[EB/OL]. https：//openai. com/index/chatgpt/.
[14] OPENAI. Gpt-4 technical report[J]. arXiv：2303. 08774，2023.
[15] WEI J，WANG X，SCHUURMANS D，et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems，2022，35：24824-24837.
[16] ZENG Z，YU J，GAO T，et al. Evaluating large language models at evaluating instruction following[C]// The Twelfth International Conference on Learning Representations. 2024.
[17] DU Y，LI S，TORRALBA A，et al. Improving factuality and reasoning in language models through multiagent debate[C]// Forty-first International Conference on Machine Learning. 2024.
[18] GU Z，ZHU X，YE H，et al. Xiezhi：An ever-updating benchmark for holistic domain knowledge evaluation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024：18099-18107.

XIONG Zhuozhi，born in 1999，postgraduate. His main research interests include knowledge graph and natural language processing.

XIAO Yanghua，born in 1980，Ph. D，professor，Ph.D supervisor，is a member of CCF（No. 12210D）. His main research interests include knowledge graph and natural language processing.

（责任编辑：柯颖）

CCF 互联网专委会、网络与数据通信专委会与 ACM SIGCOMM 合作签约仪式在沈阳举行

2025 年 9 月 13 日上午，在沈阳举办的 CCF 网络大会期间，CCF 互联网专业委员会、CCF 网络与数据通信专业委员会与 ACM SIGCOMM 举行战略合作签约仪式。CCF 理事长孙凝晖教授现场见证，CCF 互联网专委会主任王兴伟教授与 ACM SIGCOMM 主席 Matthew Caesar 教授作为双方代表完成签约。

本次战略合作聚焦四大领域：

1）国际平台共建

ACM SIGCOMM 将支持 CCF 网络大会国际化进程，推动中国网络学术成果的全球传播。

2）科研激励计划

联合设立面向中国网络领域科研人员的专项奖励机制。

3）人才联合培养

共同组织网络科研暑期学校，培育高端技术人才。

4）学术生态拓展

协作开展网络科研线上学术活动，构建常态化交流平台。

SIGCOMM 会议汇集了来自学术界和工业界的中国出版物，此前在京与 CCF 会谈时，其代表表达了希望通过举办学会活动、夏令营合作、国际网络项目、奖励等促进双边学者沟通及技术发展的愿望。CCF 和 ACM SIGCOMM 双方认为彼此之间未来可就通讯领域技术及学术进行深度交流，促进学会发展及技术进步。

据 CCF 微信公众号