

文章编号: 1003-0077(2025)06-0168-11

大语言模型的中文多轮指令遵循能力评测研究

朱 秦, 何俊亮, 邱锡鹏, 黄萱菁

(复旦大学 计算机科学技术学院, 上海 200433)

摘 要: 该文提出了一种高效评测中文大语言模型(LLM)指令遵循能力和多轮对话能力的方法,并构建了中文多轮指令遵循基准(Chinese Multiturn Instruction Following Benchmark, CMIF)。该文研究设计了专门针对中文的原子指令数据集,涵盖语言结构、拼音、音调等特性,并结合规则与 LLM 对多轮问题的合法性进行复查,确保评测结果的准确性。在实验中,选取了包括 GPT4o 和 Qwen2.5-72B-Instruct 在内的 14 个开源及闭源模型进行评估。结果显示,主流模型在单轮对话场景中具有较好的指令遵循能力,但多轮对话表现仍有较大提升空间。其中,单轮指令级准确率最高的 Claude-3.5-Sonnet 在多轮场景下准确率从 73.8% 下降至 40.0%。此外,这些模型在处理中文原子指令时表现出明显的性能下降,中文任务的综合准确率最高仅为 51.0%,显著低于其他四类任务平均 79.0% 的综合准确率。

关键词: 指令遵循; 多轮对话; 大语言模型; 数据集构建

中图分类号: TP391

文献标识码: A

Evaluation on Chinese Multiturn Instruction Following Capability of Large Language Models

ZHU Qin, HE Junliang, QIU Xipeng, HUANG Xua njing

(School of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: This study proposes an efficient method to evaluate the instruction-following and multiturn dialogue capabilities of Chinese large language model (LLM). It introduces the Chinese Multiturn Instruction Following Benchmark (CMIF). A dataset of atomic instructions tailored to Chinese is constructed, covering language structures, pinyin, and tones. Rules and LLM are employed to verify the validity of multiturn questions, ensuring accurate evaluation results. Fourteen open-source and closed-source models, including GPT4o and Qwen2.5-72B-Instruct, are evaluated. Results show that while most models perform well in single-turn scenarios, their performance in multiturn settings declines significantly. Claude-3.5-Sonnet achieves the highest single-turn instruction-level accuracy, but in multiturn scenarios, its accuracy drops from 73.8% to 40.0%. Additionally, the models demonstrate notable performance degradation when processing Chinese-specific atomic instructions. The highest overall accuracy for Chinese tasks, achieved by GPT4o, is only 51.0%, far below the 79.0% average across other task categories.

Keywords: instruction following; multi-turn dialogues; large language model; dataset construction

0 引言

大语言模型(Large Language Models, LLM),如 ChatGPT 系列^[1-2],近年来在自然语言处理(Natural Language Processing, NLP)领域取得了显著进展,

因其卓越的指令遵循能力和多轮对话能力成为研究热点。指令遵循能力使模型能准确理解并执行用户任务,生成符合语法和语义的输出,展现出高度灵活性和适应性;多轮对话能力则使其在复杂对话中保持上下文连贯性,合理推理,生成逻辑一致的回答,提供自然流畅的对话体验。这要求模型能够理解对

收稿日期: 2024-12-30 定稿日期: 2025-01-13

基金项目: 联合基金项目(U24B20181)

话历史,进行情境推理,并动态调整回答策略。

在中文领域,LLM 的指令遵循能力和多轮对话能力已经得到了广泛关注和实证研究^[3]。然而,如何高效、准确地评估这些能力仍然是一个亟待解决的问题。当前的研究中,虽然已有若干高质量的评测集和评估方法被提出,但大多数现有评测方法更侧重于单一任务的性能评价,如 IFEval 仅测试模型遵循一轮指令的准确率^[4],又如 AlignBench 仅评估模型回答一轮问题的质量,对于模型在多任务、多回合对话中的表现评价体系尚不完备^[5]。此外,现有的评测标准和方法大多局限于英文语境,对于中文的适应性和准确性尚未充分验证。

目前,大语言模型在中文语境下的指令遵循能力和多轮对话能力的评测仍处于初步探索阶段。具体来说,现有研究尚缺乏系统化和标准化的框架来全面评估 LLM 在中文环境中的表现。本文旨在填补这一研究空白,重点探讨如何设计和实现高效且准确的评测方法,以评估大语言模型在中文指令遵循与多轮对话能力方面的表现。

本文将重点解决以下两个研究问题:

(1) 如何有效评估大语言模型在中文语境下的指令遵循能力?

在大语言模型的研究中,指令遵循能力是衡量模型实际应用能力的关键维度之一。然而,现有的大多数中文指令遵循测试集多来源于英语或其他非中文语言的翻译,使得这些数据集无法充分反映中文语言环境中的实际情况。直译的数据不仅存在如规定大小写等只适合于拉丁文字的查考内容,而且往往未能充分考虑中文的语言结构、用词习惯及文化背景,难以全面评估中文特有的表达方式与限制条件。为解决这一问题,本文在前期研究中提出的高质量中文数据集基础上,设计了一类专门针对中文的原子指令,涵盖中文特有的语言结构、拼音、音调等特性。此外,现有一些评估工作依赖大语言模型(LLM)作为评分标准,然而这种方法可能导致评估结果受到强 LLM 性能的影响,从而失去准确

性^[6]。为确保评估的全面性与准确性,本文设计并采用了基于规则的评测方法,在降低评估代价的同时,力求在客观性与精确性方面提供更为可靠的评估体系。

(2) 如何高效地评估大语言模型在中文环境下的多轮对话能力?

针对这一问题,基于真实的中文用户提问数据,并结合设计的原子指令,本文提出一种自动化生成中文多轮指令遵循问题的方法。在这一过程中,LLM 需要在每一轮对话中结合当前的原子指令,对前几轮的输出进行调整和优化。为确保多轮对话的输出合理性与一致性,本文引入了规则与强 LLM 相结合的策略,用以判断多轮原子指令组合的合法性,从而保证存在输出结果能够满足所有轮次的原子指令要求。具体而言,本文从两个维度对模型的多轮对话能力进行评测:一是“对话维度”,即要求模型对话的每一轮指令均正确实现,考察模型多轮对话综合能力;二是“指令维度”,即只统计模型各原子指令的成功率,考察模型独立遵循原子指令的能力。通过这两个维度的准确率分析,本文旨在对模型的多轮对话能力进行深入的、多层次的评估。

表 1 展示了本文提出的中文多轮指令遵循基准(CMIF)与其他主流基准相比的优势。

本文的主要贡献如下:

(1) 设计了 25 类中文特色指令,覆盖大语言模型在中文应用场景下的五大常见使用场景。

(2) 基于特色指令构建并开源了 CMIF 评测集及配套评测方法,为准确评估大语言模型的中文指令遵循能力和多轮对话能力提供了标准化工具。

(3) 对主流中文大语言模型进行了系统性评测,从多个维度深入分析了模型表现,提供了细粒度的性能洞察。

(4) 通过实验结果分析,识别了现有模型在中文能力方面的关键短板,为后续模型优化指明了具体改进方向。

表 1 CMIF 和其他数据集的对比

数据集	数据量	数据来源	语言	评估指标	是否多轮	是否针对中文设计
MMLU ^[7]	15 908	考试和教材	英文	准确率	否	否
MT-Bench ^[8]	80	模型合成	英文	模型打分	是	否
GSM8k ^[9]	8 000	人类标注	英文	准确率	否	否
CMMLU ^[10]	11 528	考试和教材	中文	准确率	否	否

续表

数据集	数据量	数据来源	语言	评估指标	是否多轮	是否针对中文设计
AlignBench ^[8]	683	真实场景数据	中文	模型打分	否	否
IFEval ^[4]	541	模型合成	英文	准确率	否	否
MultiIF ^[11]	4501	模型合成	翻译	综合准确率	是	否
CMIF(本文提出)	570	真实场景数据	中文	综合准确率	是	是

1 任务定义

多轮指令遵循能力评测作为一种综合评估任务,旨在测试大语言模型在多轮对话中对指令的理解、跟进与执行能力。该任务要求模型能够在多个交互步骤中理解并适应复杂的指令,准确执行每一轮任务,并在后续对话中保持上下文一致性。例如,假设在一个对话中,用户首先询问“世界上第四长的河流是哪条河?回答字数不超过100字”。在该轮中,模型应提供一个字数符合格式的答案。然后,用户进一步要求模型在回复中添加一条字数严格相等的排比句,并将排比句用“&&”符号框起来。模型在此轮中需要保持第一轮的字数量要求,同时满足包含排比句的新要求。在这个任务中,模型需要在每一轮中理解用户的要求,并根据之前轮的指令提供准确的后续回复。最终,通过分析模型的每一轮输出与指令要求的一致性,来测试模型的多轮指令遵循能力。

2 构建与评测方法

为了具体研究如何评测 LLM 的中文多轮指令遵循能力,本节将分别展示中文指令和多轮对话测试集的组成、中文指令和多轮对话测试集构建方法,以及该测试集的评测指标与评测方法。

2.1 数据集组成

每一条 CMIF 中的样本包括一个原始问题、三条指令,其中每一个样本的每一条指令拥有一个类别,这些类别如下:

关键词 关键词任务限制模型在回答时使用特定词的频率。例如,可以要求模型在回答中使用某些特定的词汇、限制某个词出现的次数,或不允许在回答中使用某个特定词。这类任务帮助检验模型在特定语言条件下的灵活性和准确性,尤其是在需要

严格遵循语言表达要求或避免重复时。由于原数据集提供了问题的参考答案,所以我们的关键词均从参考答案中采样,以兼顾指令的合理性和难度。

长度限制 长度限制关注模型在回答中控制字数、段落数等方面的能力。通过限定回答的字数、段落数、某一段的字数范围等,长度任务旨在评估模型在遵循特定结构要求下的表现。例如,有些任务要求模型的回答在规定的字数范围内,这可以考察模型是否能精确地控制信息的呈现量。这类任务的设计常常要求模型不仅要有清晰的表达思路,还要能够精确地调整内容的分配与组织。

特殊符号 特殊符号任务更加关注模型在回答中的格式要求,包括是否使用特定符号、标点符号或排版要求。例如,模型可能被要求使用书名号来标明某个标题,或者在回答中使用加粗格式、列表等。这类任务的设计不仅考察模型的语言能力,还涉及格式和结构的正确应用。符号任务在实际应用中可能用来检验模型在特定格式下的信息组织能力,尤其是在那些需要严格格式规范的场景下,符号任务能够帮助提升答案的可读性和规范性。

额外生成 额外生成任务主要测试模型在面对用户提出的附加生成要求时的表现。例如,用户可能要求在回答问题前先复述题目,或提供多个可能的答案供用户选择。

中文特色 中文特色任务则要求模型在回答中使用特定的语言风格或形式。例如,拼音检查任务要求模型在回答的第二部分提供拼音标注,而排比句样式任务要求模型在回答中使用结构对称、形式一致的排比句。这些任务的设计主要针对模型在处理中文不同语言形式时的表现力。此外,文言文样式任务要求模型用类似古文无标点的格式表达内容,进一步测试其在处理古代文学语言时的能力。

2.2 数据集构建方法

如图1所示,构建测试集中每一条指令时,本文遵循以下步骤:

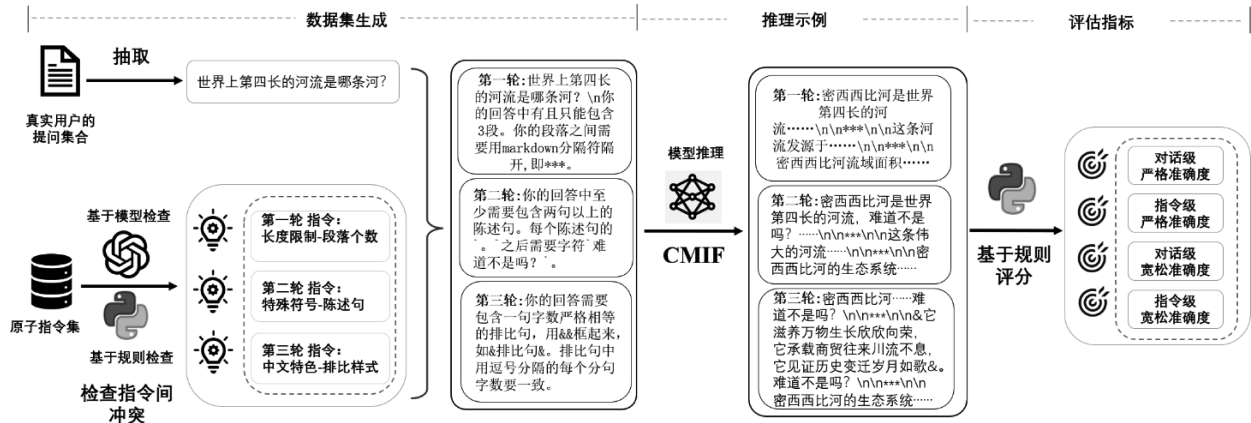


图1 CMIF 总览图

2.2.1 真实用户问题抽取

本文从选定类别的真实用户提问中抽取一个原始问题^[5]，如“世界上第四长的河流是哪条河？”。选定的类别包括综合问答、角色扮演、文本写作、中文理解、基本任务和专业能力。

2.2.2 指令抽取

本文从预先设计好的原子指令集中抽取三条指令，这些指令可能具有不同的类别。由于各指令间可能存在冲突，本文先采用了基于规则的检查方法，将一些容易识别的冲突指令去除。如关键词任务中必须使用和禁止使用同一个词等，保

证指令最基本的合理性。然而，有一些复杂的指令间冲突或者问题-指令冲突不容易通过规则来识别出，本文还设计 Prompt 采取了基于强 LLM 的模型检查，进一步识别细微的冲突，从而进一步保障指令的质量。

2.2.3 指令结合

将抽取出的一个原始问题和三条通过冲突检查的原子指令的第一条拼接，作为新指令的第一轮输入，其余两条原子指令将作为后续两轮的输入指令。至此完成了 CMIF 的一条多轮新指令的构建。最终生成的各类型指令数量分布如图 2 所示。

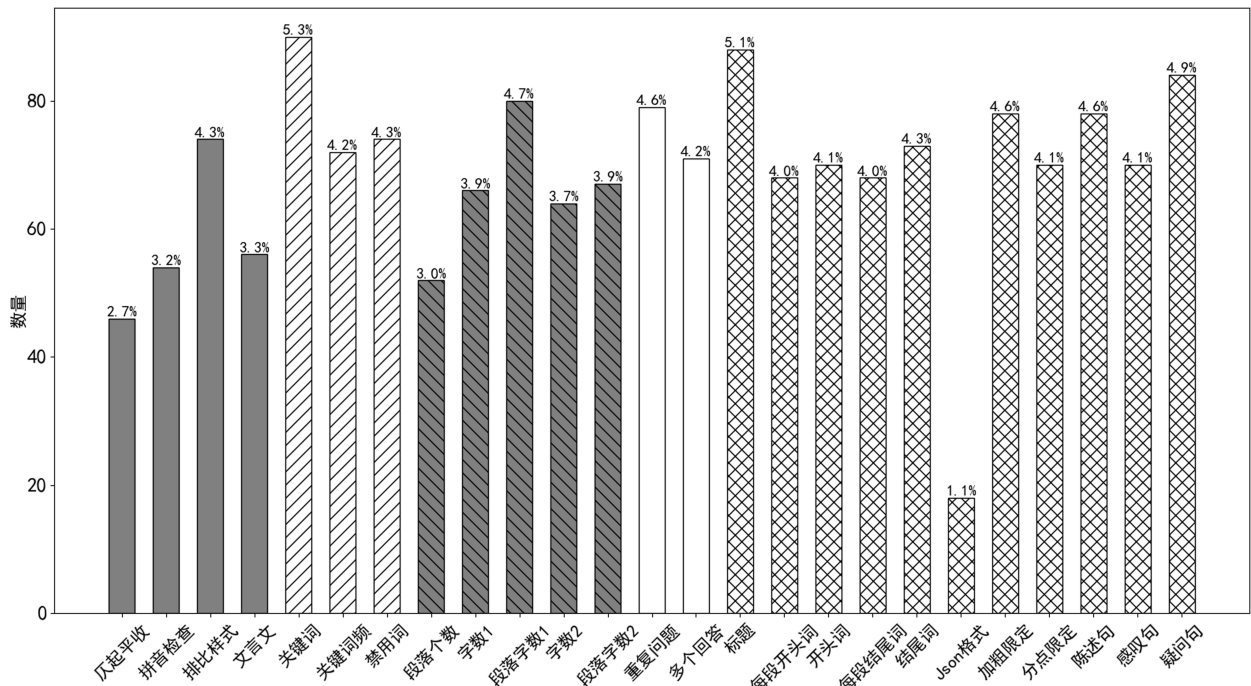


图2 CMIF 原子指令各类别占比分布图

2.3 数据集评测方法与指标

在多轮指令遵循评测中,本文采用了一个逐轮拼接指令和回复的评估方法。具体而言,输入新指令的第一轮输入,生成模型在第一轮对话中的回应,并将此回应与新的指令的第二条组合成第二轮的输入。如此反复,直到对话达到预定的轮数。在每一轮中,本文确保模型不仅要执行当前轮次的指令,还需要结合前一轮的输出,保持上下文的一致性,确保多轮对话的流畅性和逻辑连贯性。

为了全面衡量模型在多轮指令遵循中的表现,本文沿用之前工作的设定^[11],引入了四种准确率评分标准:指令级严格准确率、对话级严格准确率、指

令级宽松准确率和对话级宽松准确率。指令级严格准确率专注于每个指令的精确执行,评估模型是否严格遵循每条指令的要求。对话级严格准确率则考察整个对话过程中每一轮的指令执行情况,确保模型能够在多个轮次中始终准确理解并执行每条指令。与此对应,指令级宽松准确率和对话级宽松准确率则提供了一个更加宽松的评估标准,允许一些细微的表达差异,如引导语句或额外的非指令性内容,从而更真实地反映模型在实际使用中的表现。本文最终以指令级严格准确率、对话级严格准确率、指令级宽松准确率和对话级宽松准确率的平均值作为模型多轮指令遵循能力的评估标准,评测结果如表 2 所示。

表 2 CMIF 评测结果

模型名称	严格标准		宽松标准		综合准确率
	对话级	指令级	对话级	指令级	
开源模型					
Qwen2.5-7B-Instruct	0.140	0.492	0.170	0.539	0.335
Qwen2.5-14B-Instruct	0.170	0.555	0.209	0.597	0.382
Qwen2.5-32B-Instruct	0.218	0.592	0.258	0.637	0.426
Qwen2.5-72B-Instruct	0.281	0.651	0.337	0.692	0.490
Yi-1.5-6B-Chat	0.019	0.293	0.046	0.363	0.180
Yi-1.5-34B-Chat	0.049	0.368	0.081	0.442	0.235
InternLM2_5-7B-Chat	0.042	0.370	0.081	0.429	0.231
InternLM2_5-20B-Chat	0.081	0.427	0.137	0.496	0.285
GLM-4-9B-chat	0.091	0.458	0.126	0.507	0.296
闭源模型					
GPT4o-mini	0.202	0.595	0.263	0.650	0.428
GPT4o-0806	0.326	0.690	0.388	0.731	0.534
Doubao-pro-32k	0.282	0.657	0.361	0.707	0.502
Doubao-pro-128k	0.149	0.53	0.211	0.588	0.370
Claude-3.5-Sonnet	0.328	0.686	0.400	0.738	0.538

3 实验与分析

本节首先介绍使用的模型系列,然后通过展示实验结果和分析,来对现在模型的中文多轮指令遵循能力进行测试,并分析改进点。

3.1 模型选择

综合考虑成本和资源,本文在开源的四个系列和闭源的三个系列共计 14 个具有良好中文能力的

代表性模型上进行了实验。

开源模型 包括 Qwen2.5 系列^[12],Yi-1.5 系列^[13],InternLM2.5 系列^[14]和 GLM-4-9B-Chat^[15]。

闭源模型 包括 GPT 系列^[2],Doubao-pro 系列和 Claude-3.5-Sonnet。

3.2 主要结果

在这些模型上评测的主要结果如表 2 所示。这个结果反映出这些中文指令给模型造成了不小的挑战,大多数模型的综合准确率偏低(低于 30%),说

明这些 LLM 在中文指令遵循方面存在提升空间。少数模型展现出较好的中文多轮指令遵循能力(取得了 50%左右的综合准确率),如开源的 Qwen2.5-72B-Instruct 和闭源的 GPT4o, Doubao-pro-32k 和 Claude-3.5-Sonnet。此外,开源模型中多轮指令遵循能力最强的是 Qwen2.5-72B-Instruct,而闭源模型中是 Claude-3.5-Sonnet。

特别值得注意的是,在开源模型中, Qwen2.5-72B-Instruct 在多轮指令遵循能力方面表现最佳,而在闭源模型中, Claude-3.5-Sonnet 表现最为优异。此外,增加模型的上下文长度有可能损害模型的多轮指令遵循能力,例如, Doubao-pro-128k 的表现弱于 Doubao-pro-32k。然而,增加模型的规模(如层数、隐藏层维度)可以有效增强模型的多轮指令遵循能力,这在 Qwen2.5 系列的不同规模模型中得到了验证。

3.3 类别维度分析

本节从各类别维度分析一些表现较好的模型的综合准确率。

3.3.1 一级分类

在一级分类维度上,本文选择了 GPT-4o、Claude-3.5-Sonnet 和 Qwen2.5-72b-Instruct 作为研究对象,并对其在多个任务中的表现进行了详细对比分析,结果如图 3 所示。通过对关键词提取、长度限制、特殊符号处理以及额外生成任务等方面的系统评估,发现这三种先进的大语言模型(LLM)均表现出较为出色的能力,其综合准确率普遍超过 70%。特别是在关键词提取与长度限制任务上, Claude-3.5-Sonnet 的表现尤为突出,综合准确率接近 90%,这一结果表明该模型在多轮指令遵循、信息提取等方面具备较强的适应性和执行能力。

尽管这些模型在大多数常见任务中展现了优异的性能,但其在处理中文特定任务时的表现却相对较弱。具体而言, Qwen2.5-72B-Instruct 与 Claude-3.5-Sonnet 在中文任务中的综合准确率仅为约 40%,而 GPT-4o 的准确率约为 50%。这一趋势与所有模型在中文任务中的平均表现(约 30%)基本一致,凸显了当前大语言模型在中文语言处理方面,尤其是在指令遵循与任务理解的准确性上,仍然存在显著不足。

从分析结果来看,中文任务的较低准确率可能与以下几个因素有关:首先,中文语言具有独特的语法结构与词汇特征,这使得在中文语料上训练不充分的 LLM 在处理中文时可能面临一定的适应性

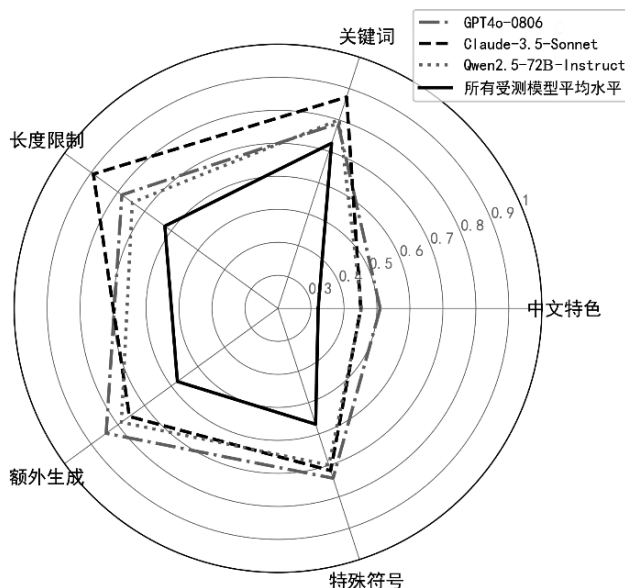


图3 基于一级分类的模型表现

问题;其次,中文任务中常常涉及文化背景知识,语境依赖性较强,这要求 LLM 在推理能力与跨语境理解能力方面具备更高的灵活性;最后,中文任务的特殊符号与结构形式可能带来语义理解上的难度。

3.3.2 二级分类

进一步地,本文对各二级分类任务中的所有模型的平均准确率进行了详细分析,结果如图 4 所示。结果显示,LLM 在关键词使用、禁用关键词、字数限制和加粗限定等任务上的表现尤为出色,平均综合准确率均达到了 90%以上,说明 LLM 在处理这些具体指令时展现出较高的准确性和一致性,能够有效地遵循任务要求,并生成符合预期的输出。

然而,另一方面,在仄起平收、拼音检查、关键词频率控制、陈述句、疑问句以及感叹句等任务中的表现则相对较差。具体而言,这些任务的平均综合准确率均低于 30%。这一现象表明,尽管 LLM 在一些结构性较强、规则明确的中文任务中表现优异,但在处理涉及更复杂语言结构和语法规则的中文任务时,模型的能力仍显不足。特别是在涉及拼音或语法细节的任务中,LLM 未能充分理解或准确执行相关要求,导致输出结果的质量不尽人意。

从模型性能的差异来看,可能存在多种潜在因素:首先,LLM 在语言生成任务中往往侧重于大的文本模式学习,对于某些特定的语法规则或音韵规律的掌握较为不够。其次,仄起平收、拼音检查等任务涉及的是中文的音韵结构和声调特性,这对于模型而言是一项相对复杂的挑战,尤其是缺乏语音输入的支持时,模型难以准确捕捉到细微的语音差异。

最后,陈述句、疑问句、感叹句等任务涉及的是语言的情感色彩和语气变化,这类任务不仅要求模型掌握语法规则,还需要其理解语境和情感信息,而现有

的 LLM 在情感推理和细粒度语义理解方面仍存在一定的局限性。

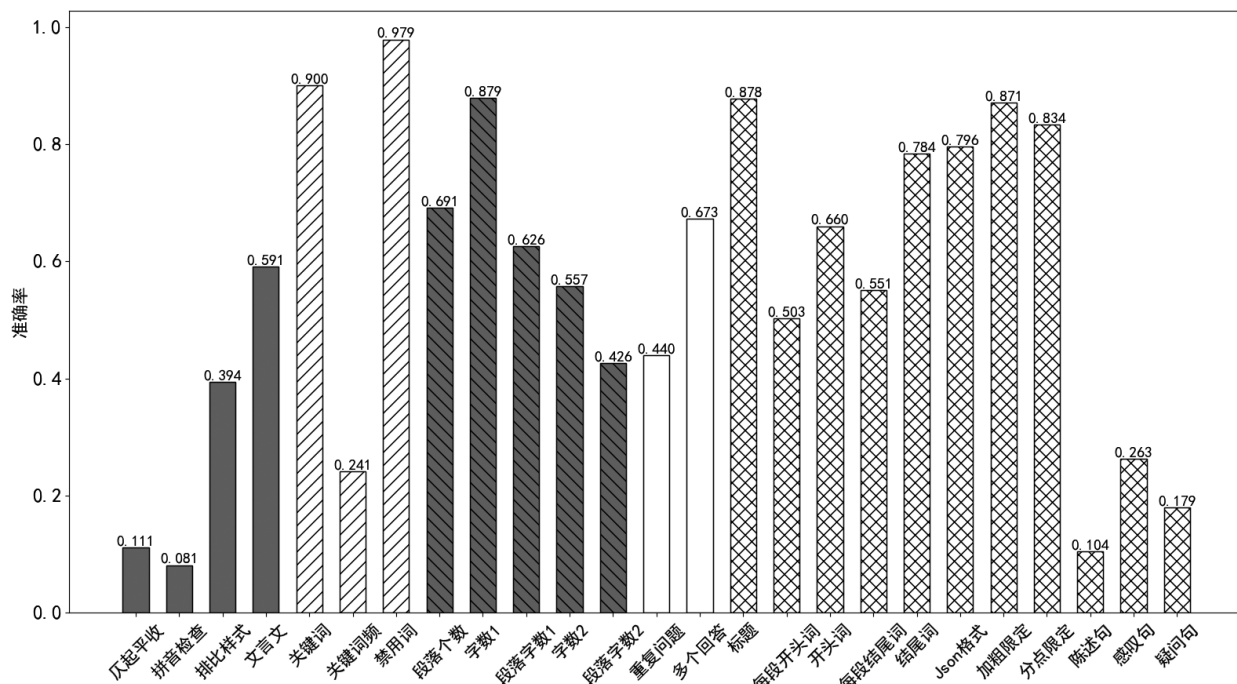


图4 基于二级分类的模型平均表现

3.4 多轮准确率分析

本节分析了 GPT-4、Claude-3.5-Sonnet、Doubao-pro-32k 以及 Qwen2.5 系列在多轮对话中的指令遵循准确率表现,如图5所示。

实验结果表明,所有模型的指令遵循准确率均随交互轮次增加呈现下降趋势。这一现象可能源于模型需要处理的信息和指令量随轮次增加而累积,导致其指令遵循能力逐渐降低。尽管当前大模型普遍具备较强的上下文理解能力,但随着输入信息的积累,模型可能会出现信息过载,影响注意力的有效

分配,从而降低指令遵循的准确性。

从 Qwen2.5 系列的对比数据可以发现,模型规模与准确率下降幅度呈现负相关:72B 版本的准确率从 71.5% 降至 59.9% 再到 49%,而 32B 版本则从 68.8% 降至 54.2% 再到 42.6%。这表明较大规模的模型在长期对话中能够保持相对稳定的指令遵循能力,可能得益于其更大的参数量和更丰富的训练数据,使其在处理多轮指令时能够更好地维持上下文一致性。然而,值得注意的是,即便是规模较大的模型,在交互轮次持续增加的情况下,其指令遵循准确率仍会出现显著下降,这反映了当前大模型在长期

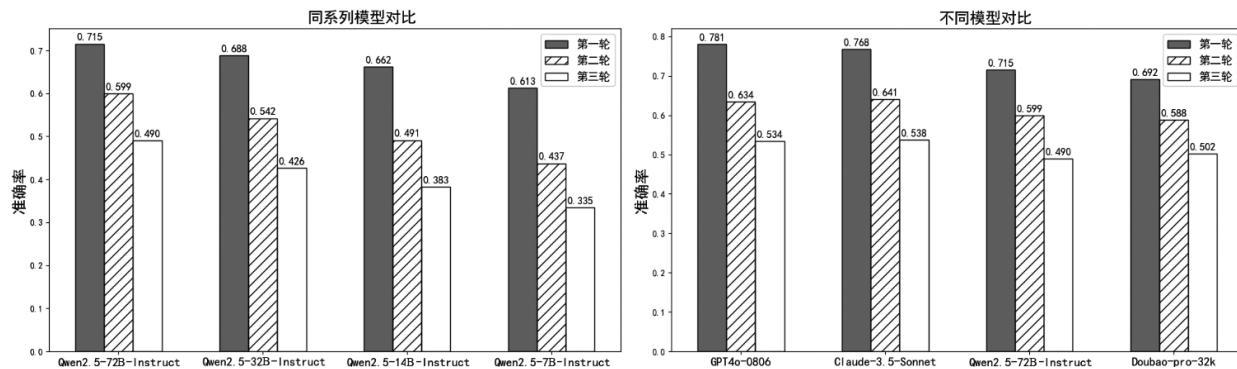


图5 多轮准确率变化图

信息处理方面仍存在普遍性的瓶颈。

3.5 遗忘率与修正率分析

本节对比了所选模型的遗忘率和修正率表现。遗忘率(Instruction Forgetting Rate, IFR)表示模型在上一轮正确遵循指令的基础上,当前轮未能遵循指令的比例,如式(1)所示。

$$IFR = P(\text{本轮未遵循} \mid \text{上一轮遵循}) \times 100\% \quad (1)$$

修正率(Instruction Correction Rate, ICR)则表示模型在上一轮未能遵循指令的情况下,本轮成

功遵循指令的比例,如式(2)所示。

$$ICR = P(\text{本轮遵循} \mid \text{上一轮未遵循}) \times 100\% \quad (2)$$

如图6所示,模型的修正率在不同规模下保持相对稳定,约为7%,这表明模型的修正能力与其参数规模之间并无明显关联。相比之下,遗忘率与模型规模呈现显著的负相关关系,在Qwen2.5系列中表现尤为明显:随着模型规模的减小,遗忘率逐步升高,分别为13.9%、19.1%和29.4%。这一现象说明较小规模的模型在处理新指令时,更容易出现遗忘先前指令的情况。

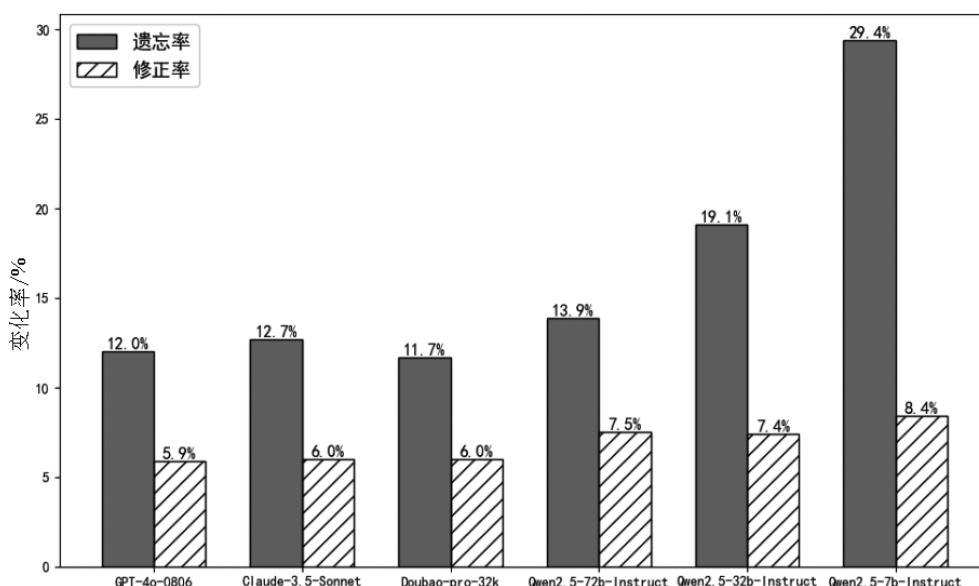


图6 遗忘率与修正率对比图

3.6 评测集可靠性分析

最后,本文通过对比实验研究,探讨CMIF与现有基准评测方法在评估模型中文指令遵循能力方面的可靠性。实验选取了Llama3.1-70B-Instruct^[16]和Qwen2.5-72B-Instruct两个代表性大模型,分别在CMIF、MultiIF和IFEval三个数据集上进行测试。

如表3所示,在英文主导的IFEval单轮评测中,Llama表现略优于Qwen(0.869 vs 0.856);而在CMIF的首轮评测中,Qwen则明显优于Llama(0.715 vs 0.627)。这种对比凸显了两个模型在中英文处理能力上的差异:Llama在英文指令方面表现更为出色,而Qwen在中文指令方面具有明显优势。这种性能差异与两个模型的训练策略存在直接关联:Meta在技术报告中明确指出,Llama的SFT

阶段仅包含少量多语种数据,主要着重于提升英文能力^[16];相比之下,Qwen团队在SFT过程中采取了中英文能力均衡发展的训练策略^[12,17]。

表3 首轮CMIF Vs IFEval

综合准确率	IFEval	CMIF(首轮)
Qwen2.5-72B-Instruct	0.856	0.715
Llama3.1-70B-Instruct	0.869	0.627

如表4所示,在MultiIF的中文部分上,Llama相较于Qwen2.5展现出优势,这与两个模型的训练策略是相悖的。然而,在CMIF数据集上,Qwen的表现明显更好,这才是更加合理的结果。这说明,尽管MultiIF包含中文评测部分,但这些内容均源自英文直译^[11],在评估模型真实中文能力时存在明显局限性,这进一步反映了CMIF数据集的可靠性。

表 4 CMIF Vs MultiIF

综合准确率	MultiIF(中文)	CMIF
Qwen2.5-72B-Instruct	0.608	0.490
Llama3.1-70B-Instruct	0.622	0.448

基于实验结果,我们得出两个关键结论:首先,仅依赖英文直译的指令评测集难以准确反映模型的中文指令遵循能力;其次,CMIF 通过专门设计的中文特色原子指令,能够更有效地区分和评估模型在中文场景下的真实表现。这也验证了开发针对性中文指令评测集的必要性。

4 相关工作

近年来,大型语言模型(LLM)在多种任务中展现出了卓越的能力^[2,12,15],其中指令遵循是最关键的能力之一,决定了模型输出与用户期望之间的契合度,在实际应用中具有重要意义^[18-20]。

4.1 指令遵循评估方法

尽管指令遵循能力至关重要,但评估这一能力仍然面临许多挑战。现有的指令遵循评估方法主要分为三类:人工评估、基于 LLM 的评估和基于规则的评估。每种方法都有其优缺点。①人工评估虽然能够较为准确地判断模型响应的质量^[1],但其高昂的成本、时间消耗以及评估者之间的个体差异,使得评估结果难以复制和标准化,影响了评估的一致性与可靠性^[21]。②基于 LLM 的评估则通过使用另一个模型来自动评估目标模型的表现^[22]。虽然这种方法比人工评估更具可扩展性、资源消耗更低,但其效果强烈依赖于评估模型本身的能力。如果评估模型存在偏差或缺陷,那么评估结果可能会出现误差,从而影响对目标模型能力的准确判断^[4]。③基于规则的评估通过预设的规则来检查模型输出是否符合指令要求,具有一定的标准化优势,但是只能评估存在客观答案的问题^[23]。

4.2 中文多轮基准

目前,大多数现有的指令遵循基准主要集中在单轮对话中,这无法充分评估 LLM 在多轮指令遵循方面的能力。然而,在现实世界的用户与 AI 交互中,大多数对话往往是多轮的,指令也会随之变

化,并且前面的指令可能会持续影响后续的结果。因此,现有的单轮评估基准显然不足以全面评估 LLM 在多轮对话中的表现。另外,其他任务中已开发出多轮基准,如跨语言的泛化能力^[24]、对话系统^[25]等。另一个值得注意的问题是,大部分指令遵循基准仅限于单一语言,通常是英语^[26],这使得这些评估方法在应对中文及其他非英语语言的应用场景时显得力不从心。清华大学的工作研究了单轮中文指令遵循数据集,并且通过模型打分进行评估^[5]。Meta 的工作通过将英文语料翻译成另外 7 种语言实现了多语言评估^[11]。然而,要确保指令遵循评估的公平性,必须考虑语言和文化差异^[27-28]。某些指令在不同语言中可能无法直接翻译,或因文化差异导致指令理解的偏差。例如,“回答全部大写”这一指令,在不同语言和文化中可能会引发不同的反应或理解方式,因此,仅通过翻译并不能准确反映模型在不同语言中的表现。基于此,本文提出了 CMIF,首个从一开始就专门针对中文进行设计的多轮指令遵循基准。

5 研究展望

本文提出了第一个针对中文进行设计的多轮指令遵循基准 CMIF 以及相应的数据集构造方法,解决了有效评估大语言模型在中文语境下的指令遵循能力和多轮对话能力两个问题。我们对 15 个国内外知名的开源和闭源模型进行评估与分析,实验观察到尽管在单轮指令遵循任务上它们有较好的表现,但是随着轮次的增加,LLM 逐渐忘记遵循之前成功执行的指令,在一定程度上导致了性能大幅下降。因此未来工作仍然需要进一步解决大模型多轮指令遗忘的问题。同时我们也观察到,强大的推理能力是存在帮助模型从前几轮的错误中恢复的潜力的,但是目前这个没有被充分地通过训练强化来应用。这些都是 LLM 未来可能的训练改进方向,指导模型的迭代。此外,在 CMIF 的基础上,中文多轮指令基准还有很多提升的空间,比如可以进一步精细化中文特色子任务,设计更多贴合用户实际需求的中文特色原子指令。此外,还能加入一些图文问题,对一些多模态任务进行考察等。总之,希望我们的基准和发现能为未来的大语言模型研究提供有价值的见解。

参考文献

- [1] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 27730-27744.
- [2] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J]. *arXiv preprint arXiv*: 2303.08774, 2023.
- [3] ZENG H. Measuring massive multitask Chinese understanding[J]. *arXiv preprint arXiv*: 2304.12986, 2023.
- [4] ZHOU J, LU T, MISHRA S, et al. Instruction-following evaluation for large language models[J]. *arXiv preprint arXiv*: 2311.07911, 2023.
- [5] LIU X, LEI X, WANG S, et al. Alignbench: Benchmarking chinese alignment of large language models[J]. *arXiv preprint arXiv*: 2311.18743, 2023.
- [6] WANG P, LI L, CHEN L, et al. Large language models are not fairevaluators[J]. *arXiv preprint arXiv*: 2305.17926, 2023.
- [7] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding [C]//*Proceedings of the International Conference on Learning Representations*, 2021: 3-7.
- [8] BAI G, LIU J, BU X, et al. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues[C]// *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024:11-16.
- [9] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[J]. *arXiv preprint arXiv*: 2110.14168, 2021.
- [10] ZENG H. Measuring massive multitask Chinese understanding[J]. *arXiv preprint arXiv*: 2304.12986, 2023.
- [11] HE Y, JIN D, WANG C, et al. Multi-IF: Benchmarking LLM on multi-turn and multilingual instructions following [J].*arXiv preprint arXiv*: 2410.15553, 2024.
- [12] YANG A, YANG B, ZHANG B, et al. Qwen2.5 technical report[J/OJ].*arXiv Preprint*, *arXiv*: 2412.15115, 2024.
- [13] YOUNG A, CHEN B, Li C, et al. Yi: Open foundation models by 01. ai[J].*arXiv preprint arXiv*: 2403.04652, 2024.
- [14] CAI Z, CAO M, CHEN H, et al. Internlm2 technical report[J].*arXiv preprint arXiv*: 2403.17297, 2024.
- [15] DU Z, QIAN Y, LIU X, et al. Glm: General language model pretraining with autoregressive blank infilling[J]. *arXiv pre-print arXiv*: 2103.10360, 2021.
- [16] DUBEY A, JAUHRI A, PANDEY A, et al. The llama 3 herd of models[J]. *arXiv preprint arXiv*: 2407.21783, 2024.
- [17] BAI J, BAI S, CHU Y, et al. Qwen technical report [J].*arXiv preprint arXiv*: 2309.16609, 2023.
- [18] ZHONG R, LEE K, ZHANG Z, et al. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections [J].*arXiv preprint arXiv*: 2104.04670, 2021.
- [19] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners [J]. *arXiv preprint arXiv*: 2109.01652, 2021.
- [20] SANH V, WEBSON A, RAFFEL C, et al. Multitask prompted training enables zero-shot task generalization [J]. *arXiv preprint arXiv*: 2110.08207, 2021.
- [21] ZHENG L, CHIANG W L, SHENG Y, et al. Judging llm-as-a-judge with mt-bench and chatbot arena[J]. *Advances in Neural Information Processing Systems*, 2023, 36: 46595-46623.
- [22] LIU Y, ITER D, XU Y, et al. G-eval: Nlg evaluation using gpt-4 with better human alignment[J]. *arXiv preprint arXiv*: 2303.16634, 2023.
- [23] HE Q, ZENG J, HUANG W, et al. Can large language models understand real-world complex instructions? [C]//*Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(16): 18188-18196.
- [24] SIDDHANT A, HU J, JOHNSON M, et al. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization[C]//*Proceedings of the International Conference on Machine Learning* 2020, 2020: 4411-4421.
- [25] BUDZIANOWSKI P, WEN T H, TSENG B H, et al. Multiwoza large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling[J]. *arXiv preprint arXiv*:1810.00278, 2018.
- [26] ZHENG L, CHIANG W L, SHENG Y, et al. Judgingl lm-as-a-judge with mt-bench and chatbot arena[J]. *Advances in Neural Information Processing Systems*, 2023, 36: 46595-46623.
- [27] BLASI D, ANASTASOPOULOS A, NEUBIG G. Systematic inequalities in language technology performance across the world's languages[J]. *arXiv preprint arXiv*: 2110.06733, 2021.
- [28] WU S, DREDZE M. Are all languages created equal in multilingual BERT? [J]. *arXiv preprint arXiv*: 2005.09093, 2020.



朱秦(2001—), 硕士, 主要研究领域为大语言模型、自然语言处理。

E-mail: zhuq22@m.fudan.edu.cn



何俊亮(2000—), 硕士, 主要研究领域为大语言模型、自然语言处理。

E-mail: jlhe22@m.fudan.edu.cn



邱锡鹏(1983—), 通信作者, 博士, 教授, 主要研究领域为大语言模型、自然语言处理。

E-mail: xpqiu@fudan.edu.cn

第二十一届全国机器翻译与多语言信息处理大会(CCMT 2025) 征文通知

第二十一届全国机器翻译与多语言信息处理大会(The 21st China Conference on Machine Translation and Multilingual Technology, CCMT2025)将于2025年9月26日至28日在甘肃兰州举行。本届会议由中国中文信息学会主办,西北民族大学承办。CCMT旨在为国内外机器翻译界同行提供一个交互平台,加强国内外同行的学术交流,召集各路专家学者针对机器翻译的理论方法、应用技术和评测活动等若干关键问题进行深入的研讨,为促进中国机器翻译事业的发展,起到积极的推动作用。

会议已连续成功召开了二十届(前十四届名为全国机器翻译研讨会CWMT,第十五届至第二十届名为全国机器翻译大会CCMT)。其中,共组织过十四次机器翻译评测,一次开源系统模块开发(2006)和两次战略研讨(2010、2012)。这些活动对于推动我国机器翻译技术的研究和开发产生了积极而深远的影响。因此,CCMT已经成为我国自然语言处理领域颇具影响力的学术活动。

除学术论文报告外,本次会议将会邀请国内外知名专家进行特邀报告,面向学生和青年学者举行专题讲座,邀请学界和产业界专家举行专题讨论会等,通过丰富多彩的形式和与会者互动探讨机器翻译最炽热的研究论点,揭示机器翻译最前沿的蓝图。同时,CCMT2025继续组织机器翻译评测,包括双语翻译(汉英、英汉、维汉、藏汉和蒙汉)、自动译后编辑(汉英、英汉)、多领域汉英机器翻译(汽车、电子、能源、经济、IT、文学、机械、医药)等多个任务。会上也会就评测工作进行学术交流和专题讨论。

本次会议论文主题包括但不限于:

- 机器翻译的词典、语料库加工技术和工具开发
- 机器翻译模型和方法,包括基于规则、实例、统计、神经网络等模型与方法
- 多语言大模型与机器翻译
- 与机器翻译相关的自然语言处理基础问题,如词法分析、句法分析、语义分析、篇章分析等

- 语言资源稀缺场景下的跨语言技术
- 机器翻译的领域自适应
- 多模态/跨模态机器翻译,可包括文本、图片、语音、视频等模态
- 少数民族语言机器翻译
- 机器翻译的前处理和后处理技术
- 机器翻译应用系统开发及应用场景设计,包括计算机辅助翻译、嵌入式翻译、交互式翻译、多语言对话、跨语言对话协同等
- 机器翻译与翻译用户相结合的方法、工具和网络服务,包括如何提高用户翻译效率、收集用户后编辑数据、后编辑服务的质量评价等
- 机器翻译评价方法
- 多语言/跨语言信息处理,包括检索、问答、信息抽取、摘要、内容生成和文本风格转换等
- 多语言内容生成与跨语言文本风格转换
- 机器翻译在其他自然语言处理任务中的应用

本次会议继续面向海内外征集原创学术论文。会议将设置亮点论文,将在与会期间进行评选,并颁发证书。此外,CCMT2025支持平行投稿。在其他会议或期刊政策允许的条件下,CCMT2025将考虑已在其上提交审核的论文,且已提交论文在CCMT2025审核期间可以另投他处。CCMT2025不考虑任何与其他会议或期刊将要发表(或已发表)的论文在内容或结果上有明显重叠的论文。提交一篇以上论文到CCMT2025的作者必须确保提交论文在内容或结果上重叠率不超过25%。

重要日期:

投稿截止日期: 2025年8月1日

论文录用通知: 2025年9月1日

论文正式稿返回: 2025年9月10日

会议日期: 2025年9月26日~28日

(中国中文信息学会)