计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：基于对话思维链指令微调的汉语写作智能评语生成研究
作者：薛嗣媛，任福继
网络首发日期：2025-03-12
引用格式：薛嗣媛，任福继．基于对话思维链指令微调的汉语写作智能评语生成研究[J/OL]．计算机工程与应用.
https://link.cnki.net/urlid/11.2127.TP.20250312.1144.006

基于对话思维链指令微调的汉语写作智能评语生成研究

薛嗣媛[1,2]，任福继[2,3,4]
1.中国社会科学院 语言研究所，北京 100732
2.国家语委中国语言智能研究中心，北京 100089
3.电子科技大学计算机科学与工程学院，成都 610054
4.电子科技大学（深圳）高等研究院，广东 深圳 518110
+ 通信作者 E-mail：xuesiyuan987@163.com

摘要：当前大语言模型在生成写作评语时普遍存在个性化不足等问题，难以针对学生写作特点和能力水平提供差异化反馈，影响了其在教育场景中的应用效果。因此，如何利用大语言模型准确识别并适配不同学生的写作能力水平，是智能教育中实现个性化教学的关键问题。为此，本研究以 Qwen、GLM、Llama 等大语言模型为基座模型，提出一种对话思维链的指令微调策略，即通过融合个体写作技巧差异从而优化大语言模型在汉语写作评语生成的效果。结果表明，微调后的大语言模型生成的评语显著优于传统两阶段训练方法和问答对话指令微调方法，在浅层语言符号和语义层面的评估上达到最优结果。另外，通过与人类教师评语比较研究发现，模型生成评语在情感度、信息量、连贯性等方面能够实现指令的有效跟随，但在正确性方面由于写作技巧的误判和生成字符限制等原因依旧与人类教师评语存在差距。本研究从语言教学视角出发，将个体差异化信息融入大语言模型训练过程，为老师“针对性教”和学生“个性化学”提供智能应用的实证参照。

关键词：自动作文评测；大语言模型；评语反馈生成
文献标志码：A　**中图分类号**:TP391.1　doi：10.3778/j.issn.1002-8331.2411-0290

Automated Feedback Generation of Chinese Essay Based on Dialogical Chain-of-Thought Instruction Fine-Tuning

XUE Siyuan[1,2], REN Fuji[2,3,4]

1. Department of Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China
2. China Language Intelligence Center, Capital Normal University, Beijing 100089, China
3.School of Computer Science and Engineering, University of Electronic Science and Technology, Chengdu 610054, China
4.The Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen, Guangdong 518110, China

Abstract：Current large language models often lack personalization when generating writing feedback, making it difficult to provide differentiated responses tailored to students' writing characteristics and skill levels. This limitation hampers their application effectiveness in educational contexts. Therefore, exploring how to leverage large language models to accurately identify and adapt to different students' writing abilities has become a critical challenge for advancing personalized teaching in intelligent education. To address this, this study proposes a dialogue chain-of-thought instruction fine-tuning strategy based on large language models such as Qwen, GLM, and Llama model. This approach integrates individual differences in writing skills to enhance the performance of large language models in generating feedback. Results demonstrate that proposed method significantly outperform traditional two-stage training methods and single-turn instruction fine-tuning approaches, achieving state-of-the-art results in the generation evaluation and semantic evaluation Furthermore, a comparative study reveals that the model-generated feedback effectively follows instructions in terms of emotional tone, information richness, and coherence while reflecting differentiation in students' writing skills. However, due to misjudgments of writing techniques and generation limitations, discrepancies in accuracy remain when compared to human feedback. From the perspective of language teaching, incorporating individualized information into the training process of large language models. It provides empirical references for intelligent applications that enable teachers to conduct "targeted teaching" and students to engage in "personalized learning".
Key words：Automated Essay Evaluation; Large Language Models; Comment Feedback Generation

基金项目：中国社会科学院青年启动计划（2025QQJH15）；国家语委项目（WT145-59，ZD145-12）；中国社会科学院语言学重点实验室（2024SYZH001）；科技部科技创新 2030“新一代人工智能”重大项目（2020AAA0109700）的研究成果之一。
作者简介：薛嗣媛(1993—)，女，博士，助理研究员，CCF 会员，研究方向为自然语言处理、大语言模型、计算机辅助语言教学；任福继(1959—)，男，博士，教授，研究方向为情感计算、自然语言处理

在语言教学中，教师往往因无法提供大量个性化写作反馈，限制了写作训练的效果。写作智能评估系统能够利用自然语言处理技术，能够准确评估语法、结构、内容和表达等多维度写作表现，并根据以上维度生成有效反馈意见。多项研究表明，系统的及时反馈能够促进学生的写作主动性，尤其在学生提交频率、修订行为、修订深度和效果方面展现了积极成效[1]。与传统教师的反馈方式相比较，接受机器反馈的学生写作训练更频繁、写作成绩更优异[2]。在数智时代，学习模式正逐步从标准化向个性化转变，从被动接受转向主动探索。然而，现有的大语言模型在生成评语时缺乏个性化，未能根据学生写作特点提供精准反馈。因此，研究者亟须探索更精准的反馈机制以满足学生的多样化需求。理想情况下，系统应准确识别并适配不同学生的写作能力水平，为他们提供量身定制的反馈。然而，传统的问答指令微调模型往往只是将输入与输出直接关联，缺乏对问题解决过程的深入推理，因此难以提供足够深入的评价与建议。

近年来，以 ChatGPT 为代表的大语言模型冲击了以教师为中心的写作教学模式，推动教育领域从基于规则匹配的传统方式转向差异化写作评语生成[3]。大语言模型通常是在大规模语料库上进行训练，具有较强的通用能力，但在特定领域的精准匹配上存在一定局限。指令微调技术为大语言模型提供了“定制化”能力，即通过在特定领域数据集上进行指令微调，大语言模型能够更准确地遵循用户指令，从而更好地满足用户使用需求[4][5]。

鉴于此，本文提出了基于对话引导的思维链指令微调策略。该策略将写作能力的差异化信息融入模型，强化模型对个体写作技巧的关注，将原本“写作文本—评语”这一单一映射扩展为由多个中间推理步骤组成的链式对话。借助思维链式引导，模型不仅需要关注生成何种评语，还需要关注从什么角度生成评语，从而提升生成内容的可解释性和个性化。这一训练策略不仅为教师和学生提供了更具针对性的反馈，也为提升大语言模型写作智能评测的性能提供了新的途径。

1 相关研究

1.1 写作评语智能反馈研究

智能评语生成作为文本生成的一项任务，跟随自然语言处理技术发展呈现出三个不同阶段的研究范式：基于模板规则匹配、基于深度学习方法、基于生成式人工智能技术。

基于模板规则匹配的方法要求研发者根据特定的应用场景和需求预先设计反馈模板。该方法的优点在于可以根据不同任务需求进行模版定制，提供标准化、结构化的反馈。然而，在实际应用中，通常依赖于预设模板匹配写作内容的智能反馈系统，往往高度依赖专家知识和教学经验，造成了巨大的人工负担。通常情况下，模板匹配方法仅在特定情境下有效，泛化性能较差，大规模推广的难度较大，难以满足多样化的教学需求和个性化的学生需求。

随着深度学习的飞速发展，逐渐被应用于写作智能评估系统。写作智能评测系统首先根据不同维度对写作文本进行评分，随后根据各维度的评分结果形成综合评价[6]。此后，序列到序列模型引入了“编码－解码”框架，能够在更广泛的上下文中表征语义信息，解决了生成模型输入输出长度不一致问题，极大程度减少了对写作评语的标注依赖，实现端对端的文本生成[7]。尽管如此，基于深度学习文本生成研究仍面临两大挑战：一是通用模型难以适应特定领域的差异化需求，二是生成内容的可控性不足、解释性不足[8]。

随着生成式人工智能的兴起，有研究者探索大语言模型在写作评估任务中的应用潜力。Mizumoto 等人探索 ChatGPT 在自动作文评分研究中的可靠性，并通过强化大语言模型中的语言特征提高评分准确性，结果表明大语言模型可以为人工评估提供部分有价值的数据支持[9]。Yancey 等人基于欧洲语言共同参考框架预测二语学习者作文水平，结果大语言模型未能超越 XGBoost 基线模型或者人工注释者所取得的分数[10]。薛嗣媛等以汉语二语学习者为研究对象，采用标准提示、思维链提示以及自洽思维链提示等不同提示策略验证大语言模型在写作自动评分和自动评语反馈方面的性能[11]。Naismith 等人训练 GPT-4 以人类专家评估者一致的方式评估篇章连贯性，结果发现 GPT-4 在连贯性评价方面与人类评分相当，并且能够生成有效回复[12]。Meryer 等人让学生根据 GPT-3.5-turbo 生成的反馈展开写作修改训练，探讨了使用大语言模型生成反馈对提升高中生英语写作效果的影响[1]。

众多研究结果显示，虽然大语言模型可以作为写作评估的辅助工具，但在实际应用中仍需结合人工反馈和修改。

1.2 指令微调相关研究

指令微调是通过由“指令－输出”对组成的微调数据集进一步优化大语言模型的训练过程。通过指令微调的大语言模型能够更好遵循人类指令，增强模型对特定任务的理解和应用能力[13]。不同类型的指令交互模式对模型性能提升起关键作用。指令微调相关策略包括问答指令微调、思维链指令微调、多轮指令微调等。问答指令数据通常涉及模型对单一任务或问题的直接响应，评语内容往往具有较高的通用性，但在提供具体、建设性意见方面可能稍显不足。思维链指令微调一种通过逐步展示推理过程来增强模型推理能力的方法。对话指令微调则通过在连续交互中增强大语言模型对任务的理解能力。多轮对话指令调优的目标是使大语言模型通过处理复杂的、动态变化的连续对话场景，从而具备增量学习的能力。Xu 等人利用 ChatGPT 创建对话，以自问自答方式收集多轮指令数

据 Baize[14]。Vicuna 使用来自 ShareGPT 平台的日志来细化指令[15]。在推理任务上，Wei 等人提出思维链提示策略，显著提升了大语言模型在各类自然语言处理任务上的推理能力[16]。此后学术界产生了一系列思维链推理工作，如 STaR 利用大语言模型自身生成的理由进行自主改进[17]。SpecialFT 将大语言模型用作教师模型，并利用知识蒸馏技术将推理能力从大语言模型转移到小语言模型[18]。Orca 从大语言模型生成的丰富信息中模仿大语言模型推理过程，包括解释轨迹、逐步思考过程和其他复杂指令[19]。然而，尽管不同指令微调策略在提升模型性能方面表现出色，但在中文领域的应用仍然面临挑战。原因之一是特定任务中文指令数据的稀缺性使大语言模型在执行以中文为主的任务时容易受到限制。因此，面向特定任务构建高质量中文指令微调数据集尤为重要。

综上，本文将探索如何利用指令微调策略增强大语言模型在推理过程中对个体差异的理解，提升机器生成评语的建设性，促进写作智能系统在教学场景的适用性。

2 研究方法

本实验整体框架如图 1 所示，分为对话指令数据集构建、基于对话思维链引导的指令微调、评估生成结果的全面评估三个步骤。研究方法部分主要讨论写作评语智能模型构建的问题定义、数据集构建以及指令数据模型微调训练。

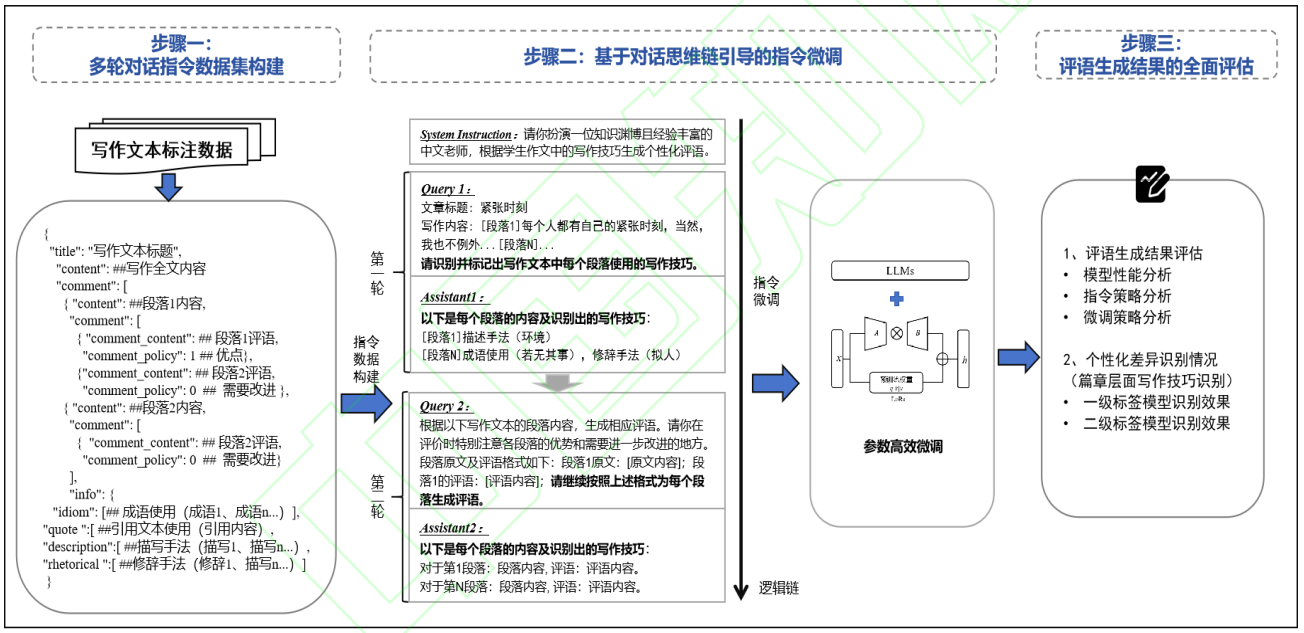

图 1 总体框架

Fig.1 The Overall Framework

2.1 问题定义

本研究旨在微调大语言模型生成针对写作文本的个性化评价。写作智能评语生成以写作文本为输入，并输出该文本的个性化评语。该问题可形式化定义为：

$$Y = argmax\ P(Y \mid X, W; \boldsymbol{\theta}) \quad (1)$$

其中，Y 表示机器生成的最优评价，X 为输入的写作文本，W 是与辅助写作评估的相关信息，$P(Y|X,W;\theta)$ 表示模型在给定写作文本 X 和辅助信息 W 的情况下，生成写作评语 Y 时的概率分布，θ为模型参数。

2.2 对话指令数据集构建

根据图 1 中步骤一所示，原始标注数据集包括学生写作文本 X，以及每篇写作文本对应的写作技巧 W 和写作评语 Y。问答指令微调构建思路是构建“写作文本，写作评语”数据。对话思维链指令微调数据构建思路是将“写作文本，写作技巧，写作评语”信息拆分成多轮指令问答，即让模型识别写作文本中的写作技巧，然后基于这些写作生成相应写作评语内容。给定原始数据$\{(X, W, Y)\} \in D$将其转换为两轮对话形式：

$$q^{(1)} = prompt^{(1)} \,||\, X \quad (2)$$

$$a^{(1)} = \{(X_1, W_1), (X_2, W_2), \ldots, (X_n, W_n)\} \quad (3)$$

$$q^{(2)} = prompt^{(2)} \,||\, a^{(1)} \quad (4)$$

$$a^{(2)} = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\} \quad (5)$$

其中，第一轮问答（$q^{(1)}$，$a^{(1)}$）聚焦于识别写作文本中哪些段落使用了写作技巧。$q^{(1)}$为第一轮对话的提问，由询问写作文本中所使用的写作技巧的提示语 $prompt^{(1)}$和写作文本 X 拼接构成；$a^{(1)}$为第一轮问题

的回复，明确指出写作段落及其对应的写作技巧；第二轮问答（$q^{(2)}$，$a^{(2)}$）基于第一轮的回答进一步生成评语；第二轮中，$q^{(2)}$将包含写作技巧的段落X_i和对应的写作技巧W_i输出，并通过提示语$prompt^{(2)}$引导模型参考第一轮回复，衡量写作技巧的优缺点生成评语Y_i；最后，所有样本构成了可用于多轮指令微调的数据集：

$$D_{sft} = \{(q^{(1)}, a^{(1)}), (q^{(2)}, a^{(2)})_{(X,W,Y)\in D} \quad (6)$$

指令微调模型可以在每个轮次t提出不同问题$q^{(t)}$并获取模型回答$a^{(t)}$以逐步实现对写作评语的分析推理。图 1 显示了对话指令微调样例，微调数据将采用 Alpaca 微调格式[4]，确保每个样本都包括指令、用户输入和模型输出、历史回复，具体样例见表 1。

表 1 多轮对话指令微调格式

Table 1 The format of multi-turn dialogue training data

```
{
    "instruction": "人类指令",
    "input": "写作文本",
    "output": "写作评语",  #生成具有建设性意见的评语
    "history": [
        ["上一轮指令", "上一轮回答"]  #识别写作文本中的写作技巧
    ]
}
```

2.3 对话思维链指令微调

有监督微调是指在预训练模型基础上，利用特定领域标注数据集进行训练，使模型能够适应特定任务的过程。因此，本节将结合指令微调数据集D_{sft}对大语言模型进行微调，写作评语生成概率表示如下：

$$P\left(\{a^{(t)}\}_{t=1}^{T}\right) = \prod_{t=1}^{T} P\left(a^{(t)} | a^{(:t-1)}, q^{(:t)}; \theta\right) \quad (7)$$

其中，$q^{(t)}$和$a^{(t)}$代表第 t 轮对话；模型在每轮生成回答$a^{(t)}$时会根据前序回答$a^{(:t-1)}$和当前问题$q^{(t)}$预测概率分布。为优化模型生成回答与真实回答的匹配度，采用如下负对数似然损失：

$$Loss = -\sum_{t=1}^{T} \log P\left(a^{(t)} \mid a^{(:t-1)}, q^{(:t)}; \theta\right) \quad (8)$$

损失函数将逐步调整模型参数θ使得生成的回答分布逐渐逼近人类教师真实回答的概率分布，其中对数运算 log 采用自然对数（ln，即 log 以 e 为底）简化梯度计算。

3 实验设置

3.1 实验数据集

实验所使用的数据集来源于网络教育平台，包含学生真实的记叙文写作样本，并配有专业教师的评语信息[20]。该数据集涵盖多维统计数据，包含篇章级嵌套结构的写作技巧标签。嵌套结构标签细粒度地反映了学生在写作中所使用的不同技巧，为个性化写作评语生成提供了坚实的数据支持。写作技巧的一级标签包括描述手法、成语使用、修辞手法、俗语使用、引用使用等；每个一级标签下又进一步细分为二级标签，描述手法包含味觉描写、心理描写、嗅觉描写、外貌描写、环境描写、神态描写、语言描写、动作描写、视觉描写、触觉描写。修辞手法包含拟人、排比、反问、设问、类比。表 2 为写作数据集的相关统计信息。

表 2 写作数据统计信息

Table 2 The Statistics of essay data

统计项	训练集	验证集	测试集
段落样本数	18,100	2,263	2,036
文本篇章数	3,996	540	887
标题平均长度	5.82	6.06	6.10
文本平均长度	408.89	382.85	345.08
段落平均数量	4.92	4.46	4.28
段落平均长度	97.58	96.15	79.64
评语平均长度	33.92	39.09	38.54
优点评语占比	84.38	80.69	79.47
改进意见占比	15.62	19.31	20.53

3.2 实验参数设置

本研究采用了两种高效的微调策略来训练模型，分别是冷冻参数微调和 Lora 微调，均基于 Llama Factory 框架实现[21]。所有模型训练均在 6 块 A-800 80GB NVIDIA GPU 上进行，使用的 Python 版本为 3.10.6，CUDA 版本为 12.4。在冻结参数微调实验设置中，使用 bp16 的精度，迭代次数为 3，批量大小为 4，最大序列长度 1024，学习率为 3e-4，使用 AdamW 优化器配合余弦调度器调整学习率。另外，在 Lora 微调实验设置中，秩（Rank）参数设置为 8, 缩放系数（Alpha）参数设置为 16。

3.3 模型对比选取

本研究对比的基线方法包括两大类别：经典的小型预训练语言模型，以及大语言模型的零样本推理和有监督微调方法。

首先，小型预训练语言模型是生成类任务中的经典模型：

（1）LongLM[22]：该模型首先在一个大型中文小说数据集上进行了预训练，采用的是编码器－解码器

序列生成方式实现的评语生成，对长文本生成和理解进行了系统性优化。

（2）ACG[20]：该模型基于 GPT-2 模型采用两阶段训练方法实现写作评语生成。该模型首先生成出一系列关键词，然后将关键词扩展成完整评论。

其次，随着大语言模型的快速发展，零样本推理和有监督微调已成为当前生成类任务的主流范式。本研究进一步引入了 2023-2024 年发布的多个中文开源大语言模型，以全面评估不同方法在评语生成任务中的表现，包括：

（3）Qwen-1 系列模型[23]：由阿里云发布的系列开源大语言模型，选取 Qwen1-7B-Chat、Qwen1-14B-Chat、Qwen-72B-Int4 用作零样本推理和微调训练。

（4）Qwen-2[24]：Qwen-2 继承了 Qwen1 模型优势，使用了更大规模和多样化的数据集进行优化，选取 Qwen2-7B-Instruct 模型用作零样本推理和指令微调训练。

（5）GLM-4-9B[25]：是由清华大学 KEG 实验室开发的一款通用的中英文对话模型。此模型在中文语义理解中具有较强的能力，能够胜任汉语的生成和理解任务，选取 GLM-4-9B-Chat 模型用作零样本推理和指令微调训练。

（6）Llama3-8B-Chinese-Chat[26]：是在 Llama 大语言模型的基础上扩展中文词汇训练而成的，中文词汇扩展有效提升了 Llama 大语言模型中文语义理解能力。

3.4 评语生成指标

3.4.1 指标评估

本文在测试集上使用的评价指标为：

（1）BLEU 系列指标[27]：衡量生成文本与真实文本之间的 N-gram 重叠。BLEU 局限性在于其过度依赖字符之间的匹配，容易忽视语义的准确性。其中，本研究选择 1-gram、2-gram、3-gram、4-gram 作为评语指标。

（2）ROUGE 系列指标[28]：衡量机器翻译输出与人类翻译文本之间的一致性。这些指标主要通过比较生成文本和参考文本之间的 N-gram 重叠、单词在句子中的顺序、单词匹配度进行评分。本研究使用 ROUGH-1、ROUGH-2、ROUGH-L 指标。

（3）Distinct-N 系列指标[29]：衡量生成文本多样性的指标，更高的 Distinct 分数表明模型生成了更多样化的语言。本研究使用 Distinct-3、Distinct-4 指标。

（4）BERTScore 指标[30]：使用预训练语言模型比较生成文本与参考文本中词汇的嵌入向量，通过计算词向量之间的余弦相似度衡量它们的语义重合程度。

3.4.2 人工评估

有研究者反映传统生成文本的评估指标无法充分体现大语言模型的实际效能[31][32]。本文将进一步利用人工方式评估其实际效果。

评估维度方面，Wei 等人主要从实用性和接受度两方面评估大语言模型生成内容，并通过输入胜利、平局、失败三类标签标注机器生成结果[16]。Chiang 等人从语法正确性、文本连贯度、故事可读性、主题相关度来衡量大语言模型在故事续写任务上的能力[31]。Zhang 等人从正确性、信息量、连贯性考察机器生成评语效果[20]。因此，本研究借鉴 Zhang 研究中所使用的评价维度，额外补充了教育场景中较为重要的情感度指标。正确性是指评语是否能够正确反映文本中的优缺点；信息量是指评语中是否包含有价值写作技巧信息，并基于此提供有效改进建议；连贯性是指评语的语法准确性和句子间逻辑的一致性；情感度是指评语能够符合教育场景下对学生引导的预期，展现出支持性和鼓励性；以上这些维度为评语生成质量的评价提供了全面的参考框架。

评估方式方面，本文选取 100 篇测试集写作文本，采用人工评分方式，对比微调模型生成评语和人类教师评语之间、微调前后评语之间的差异。人工评估由三位具有语文教育背景的专业人员依据各评估维度对机器生成评语与人类教师评语进行胜利、平局、失败的判定。在整个评估过程中，所有评估者均在未知评语来源的情况下进行独立判断。在评估多个评价者对同一对象进行分类或评分时的一致性时，使用 Fleiss's kappa 系数[33]衡量标注者之间的一致性。

4 实验结果分析

4.1 模型性能比较

表 3 是不同大语言模型在三种训练设置下（对话思维链指令微调、单轮问答指令微调、零样本评语生成）的评估结果。结果显示，对话思维链指令微调显著优于传统小模型的两阶段训练方法，在各项指标上达到了最佳水平，证明本研究提出的训练策略在提升大语言模型评语生成质量方面具有促进作用。模型大小方面，参数量超过 10B 的模型生成效果优于 10B 以下模型。其中，基于对话思维链指令微调的 Qwen1-72B 模型表现最为出色，在所有评估指标上均达到了最佳结果。然而，本地部署 10B 以上模型对本地算力的要求较高，限制了其在资源受限环境中的应用。鉴于此，本研究进一步考察在参数量低于 10B 模型的生成效果。Qwen2-7B 和 GLM4-9B 生成评语性能全面超越了 Qwen1-14B，并且仅以微小差距接近 Qwen1-72B 的结果。这一发现表明，小参数量的模型中通过精细优化和指令微调也可以实现接近超大模型的性能，从而在算力有限的场景下提供了一种更优的解决方案。在文本生成质量和训练推理效率上，Qwen2-7B 微调模型结果最优，实现了两者的最佳平衡。

表 3 大语言模型在评语生成任务中的结果

Table 3 The evaluation results of large language models in the task of generating comments

训练策略	模型选择	B-1	B-2	B-3	B-4	R-1	R-2	R-L	D-3	D-4	BertS
预训练语言模型	LongLM	33.40	26.16	22.98	21.13	\	\	\	6.05	7.39	\
	ACG	36.16	28.32	24.87	22.86	\	\	\	12.25	16.90	\
零样本推理测试	Qwen1-7B	24.40	18.01	12.24	14.01	32.88	19.44	26.65	65.07	71.50	96.92
	Qwen2-7B	27.83	13.09	7.11	4.48	30.19	7.85	19.47	91.61	96.12	96.79
	Llama3-8B	37.48	25.87	21.21	18.96	41.68	22.34	32.34	91.82	95.83	97.83
	GLM4-9B	42.75	36.39	33.58	31.99	55.24	41.72	45.24	85.15	89.78	97.31
	Qwen1-14B	42.76	36.64	33.66	32.10	55.32	42.22	45.73	82.37	87.20	96.91
	Qwen1-72B	43.36	37.05	34.24	32.65	55.47	42.35	45.91	84.33	88.87	96.95
单轮问答指令微调	Qwen1-7B	76.74	73.56	71.83	70.66	80.48	74.92	78.22	91.69	94.79	98.44
	Qwen2-7B	77.63	74.58	72.92	71.80	80.90	75.47	78.96	93.19	95.84	98.41
	Llama3-8B	75.63	72.07	70.21	68.97	78.33	72.26	76.05	93.03	95.82	98.40
	GLM4-9B	76.73	73.56	71.85	70.69	80.97	74.39	77.77	91.82	94.60	98.46
	Qwen1-14B	69.81	66.41	64.65	63.48	75.36	69.03	72.20	89.65	92.47	98.01
	Qwen1-72B	77.23	74.25	72.64	71.53	81.24	75.83	78.91	92.96	95.8	99.48
对话思维链指令微调	Qwen1-7B	77.42	74.23	72.71	71.55	80.99	75.38	78.86	92.85	95.64	99.48
	Qwen2-7B	77.94	74.76	73.04	71.88	81.12	75.55	78.93	93.09	95.91	99.35
	Llama3-8B	75.66	72.18	70.40	69.22	78.39	72.58	76.35	93.28	95.98	99.40
	GLM4-9B	77.77	74.70	73.06	71.94	80.86	75.49	78.89	93.12	95.91	99.40
	Qwen1-14B	77.59	74.41	72.71	71.55	80.99	75.38	78.86	92.85	95.64	99.48
	Qwen1-72B	**78.45**	**75.29**	**73.59**	**72.44**	**81.27**	**75.75**	**79.16**	**93.23**	**96.08**	**99.51**

4.2 提示策略选择

为进一步验证对话思维链指令微调策略的有效性，本研究展开对比分析。当移除多轮对话思维链微调策略，生成评语的各项指标整体平均下降 1.45%，反映了单轮指令微调模型对个性化特征捕捉不足，这也同时表明引入差异化写作文本特征能够提升模型生成评语的效果，有效增强模型对上下文的理解。其次当完全移除微调策略，仅依靠模型本身性能进行零样本评语生成时，各项指标均出现了进一步下降。与对话思维链微调策略相比，基于零样本推理生成的评语整体指标的平均值下降了 32%。在 10B 以下模型中，Qwen1 和 Qwen2 在零样本推理中未能超越传统两阶段模型的表现，Llama3 在 BLEU2、BLEU3、BLEU4 评估指标上也未超过传统小模型结果。在 10B 以上模型中，Qwen1-72B 的生成效果整体优于 Qwen1-14B，更高的参数量并未带来显著优势。

在参数量低于 10B 的模型中，Qwen2-7B 和 GLM4-9B 的表现全面超越了 Qwen1-14B 模型，并且仅以微小差距接近 Qwen1-72B 的结果。这表明，在零样本场景下，缺乏微调支持的模型在文本理解和生成能力方面存在明显不足。微调策略不仅能够显著提升模型对写作文本特征的捕捉能力，还在评语生成的准确性、一致性和针对性方面发挥了作用。

4.3 人工评估结果

表 4 是本研究提出的模型与未微调模型、单轮微调模型、人类教师进行比较时胜利、失败或平局的百分比。其中，胜利部分使用黑体标注。括号内 Fleiss’s kappa 系数值在 0.6<k<0.8 范围内表示评估者之间具有中高等一致性[33]。平均胜利率将为每个维度分配相同的权重比值，通过加权计算得出各对比项的平均胜利率、平局率、失败率。

表 4 模型生成评语的人工评价

Table 4 Human Evaluation of generated comments

评语对比来源	正确性 （k）	信息量（k）	连贯性（k）	情感度（k）	加权平均值
	胜利/平局/失败	胜利/平局/失败	胜利/平局/失败	胜利/平局/失败	胜利/平局/失败
多轮微调 Vs.人类教师	36/19/45（0.78）	**47**/36/17（0.87）	35/**37**/28（0.74）	**40**/29/31（0.81）	**39.50**/30.25/30.25
多轮微调 Vs.单轮微调	37/40/23（0.76）	**41**/39/20（0.81）	36/38/26（0.73）	**39**/35/26（0.70）	**38.25**/38.00/23.75
多轮微调 Vs.未微调模型	**47**/31/22（0.79）	**41**/27/32（0.82）	**38**/37/25（0.71）	**47**/34/19（0.85）	**43.25**/32.25/24.50

结果显示，首先，对话思维链微调模型与人类教师相比，在信息量、连贯性、情感度等维度上差异较小，证明机器生成评语能够有效跟随指令模仿人类教师生成有效评语。在正确性方面，人类教师在正确性评估中胜率较高，机器生成评语在正确性方面与人类还是存在差距，人类教师在内容的准确性上更具有优势。根据具体案例分析发现，正确性胜率低主要体现在两个方面，一个是大语言模型跟随指令进行写作技巧分析时有时会存在误判，而且大语言模型会自主增加标注人员未标注的写作技巧；另一个是大语言模型由于字符输出限制会出现评语截断的现象，以上情况在本实验中均标注为不正确。其次，对话思维链微调模型在信息量和情感度上都显著优于单轮微调模型。在微调过程中加入对个体写作技巧的识别，能够显著反映学生写作技巧方面的优缺点，提供更具有建设性的反馈。最后，与未微调模型相比，对话思维链微调模型在正确率、信息量、情感度方面都显著优于未微调模型。未微调模型评语内容虽然涵盖了写作的基本情况，但内容较为冗余，建设性不足。微调后的评语能够紧密围绕每个段落中的写作技巧点展开分析形成个性化意见。如表 5 所示，微调后模型能够识别文章的层次与逻辑，指出文章开头是否吸引人、结尾是否总结到位，并提出清晰的改进建议。在情感度方面，未微调模型表述较为中性，生成评语缺乏鲜明的情感色彩，较难引起学生的情感共鸣。微调后的评语学习了人类教师的情感表达方式和语言风格，例如使用“宝贝”“点赞”“非常不错哦！”等，使评语更具感染力。

表 5　机器评语与人工评语的表现差异

Table 5　Performance comparison between model-generated feedback and human feedback

写作文本	微调前	微调后	人工评语
Chatbot 主题：紧张时刻 文章内容：每个人都有自己的紧张时刻，当然，我也不例外，而我的紧张时刻是在老师发试卷的时候。这天，我早早地来到教室，窗外格外刺眼的照射着，使我不禁头昏脑胀，十分难受。不一会儿，一阵风吹来，窗外的树叶相互交叉着，发出十分难听的噪声，不由的使人心烦意乱。 老师终于来了，我连忙低下头，装作若无其事的样子，心一直砰砰地跳着，似乎要跳出嗓子眼了。就这样，老师开始了“坚难“的报分数，我双手紧握，双眼紧闭，脑海中不断浮现出妈妈那“竹笋炒肉丝“的身影。我还在口中不停默念：这次没复习好，我下次一定会努力，上天啊，你就放过我吧！不知不觉中，老师就突然报到了我：“雷英灿，100 分！哦一下子愣住了，长舒一口气拿到卷子又蹦又跳。 再扭头看，窗外的景色不知何时变得如此优美了，树叶吹着轻快的口笛，令人心旷神怡。暖暖的阳光洒在大地上；洒在书桌上，也洒在了我的心上....	评语：文章以紧张时刻为主题，生动地描绘了作者在老师发试卷前的紧张心情。通过细致入微的心理描写和环境描写，展现了作者在紧张时刻的内心感受。文章语言流畅，层次清晰，从紧张到放松的过程描绘得淋漓尽致。结尾处，窗外景色的变化与作者心情的转变相得益彰，富有诗意。作者能够通过文字将自己的情感真实地传达给读者，显示出较强的写作功底。**建议**在写作中多尝试不同的主题和表达方式，以丰富自己的写作技巧。	**第 1 段落**：每个人...（省略）老师发试卷的时候。 **评语**：宝贝，你的文章背景选材很好，通过学习中的小事感受到了自己的成长，选材很好，给你点赞呀！ **第 2 段落**：这天，我早早地来到教室...十分难听的噪声。 **评语**：开头非常棒，运用了先问其声的描写手法，不仅画面感强同时还能第一时间吸引到读者的阅读兴趣，从而为你的下文叙述奠定了基础，特别好！ **第 3 段落**：老师终于来了...似乎要跳出嗓子眼了。 **评语**：运用心理描写、神态描写的手法，可以恰如其分地表现出自己的心理感受，让读者感受你的成长哦！ **第 4 段落**：再扭头看...令人心旷神怡。 **评语**：宝贝，我们作文的部分可以点明“那一刻，我长大了”这个主题哦，再表达出自己的情感和懂得的道理就更好了！	**第 1 段落**：这天，我早早地来到教室...十分难听的噪声。 **评语**：描写耳边传来的声音，使文章更加立体。 **第 2 段落**：老师终于来了...要跳出嗓子眼了。 **评语**：巧用神态描写，心理描写和动作描写，使表达生动形象。 **第 3 段落**：再扭头看...令人心旷神怡。 **评语**：运用环境描写、拟人，画面感增强。

4.4 微调策略选择

本文将进一步对比不同微调策略对文本生成质量的影响。全参数微调使用预训练模型作为初始化权重，在特定数据集上继续训练更新全部参数，能够高效适应领域任务，但具有资源消耗量大、训练时间长、知识遗忘性大的特性[34]。由于本地部署的算力限制，本文分别采用冻结参数（Freeze Parameter）微调方法和低秩适配（Low-Rank Adaptation, LoRA）[35]高效参数微调方法进行模型训练。冻结参数微调方法在生成任务中将冻结所有参数，仅对特定任务的输出层或部分层进行微调，从而达到减少训练所需的计算资源。LoRA（Low-Rank Adaptation）高效参数微调方法通过在模型的参数矩阵上应用低秩分解，减少了参数训练量和计算复杂度，从而确保在有限算力条件下实现高效微调。

图 2 显示，LoRA 高效微调策略在语义一致性、流畅性、多样性等评价指标上表现更为出色。与冻结微调相比，LoRA 微调在不显著增加计算复杂度的情况下，允许模型进行更精细的调整，更灵活地适应评语生成任务。同时，在固定其他参数设置的前提下，冻结微调策略比 LoRA 微调策略额外耗时约 18.70 分钟，显示传统冻结参数微调虽然在一定程度上限制了模型参数的更新范围，但是模型仍保留了较大的计算结构，从而在推理阶段增加了前向传播的计算量，推理效率较低。

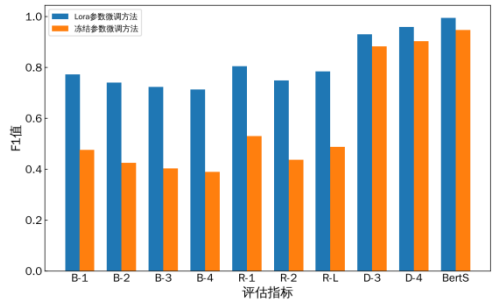

图 2 不同微调策略的结果评估

Fig.2 Evaluation results of different fine-tuning strategies

4.5 写作技巧识别

写作技巧识别是评语生成模型体现个体差异的主要途径。表 6 表示不同模型在写作技巧判别任务中存在差异。微调 GLM4-9B 模型能够有效识别嵌套结构的写作技巧，在平衡类别的条件下宏观 F1 值达到 90.8%。Qwen2-7B 在多数指标上的表现仅次于 GLM4-9B，F1 值达到 87.9%。Qwen1 系列模型整体表现不佳，尤其是 Qwen1-7B 在所有模型中性能最低，由于模型参数较少导致模型对复杂语义结构的理解能力不足。Qwen1-14B 的 F1 值比 Qwen1-7B 高出 1.8%。Qwen1-72B 的识别效果仅仅接近于 Qwen2-7B。

表 6 基于大语言模型的写作技巧识别结果

Table 6 The writing skill detection with fine-tuned LLMs

Model	Qwen1-7B			Llama3-8B			GLM4-9B			Qwen2-7B			Qwen1-14B			Qwen1-72B		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
微观	71.2	64.7	53.1	71.0	66.9	68.7	87.4	89.2	**88.3**	82.0	82.6	82.2	68.7	58.6	62.6	82.0	82.0	82.6
宏观	56.7	72.8	68.7	85.8	85.3	85.6	90.6	91.0	**90.8**	87.9	88.0	87.9	74.5	70.1	72.2	82.2	87.9	88.0
汉明损失	0.136			0.055			0.043			0.056			0.129			0.056		

图 3 是各个大语言模型在一级写作技巧标签上的加权 F1 值。结果显示，所有模型都在成语使用和引文使用达到了较好的识别效果。在修辞手法识别方面，GLM4-9B 表现最佳，F1 值达到 98.52%，而 Qwen1-7B 在这一维度的识别效果较差，F1 值为 85.79%。描写手法识别效果在所有模型中相对较弱，GLM4-9B 的 F1 值为 78.30%，而 Qwen1-7B 的识别率仅为 50.72%。

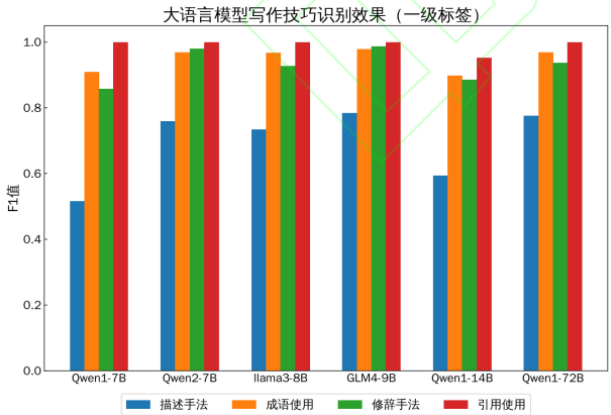

图 3 大语言模型在写作技巧识别任务中的表现

Fig.3 The writing skill detection performance of LLMs

汉明损失是用于衡量模型预测与真实标签之间的一种误差值，指标越低表明模型识别效果越好。图 4 显示各个大语言模型在一级写作标签方面的汉明损失。各类大语言模型在描述手法和修辞手法方面误差率较高，描写手法的汉明损失达到 0.67，修辞手法的汉明损失达到 0.52。由此可见，描写手法和修辞手法是各个大语言模型表现存在差异的主要原因。

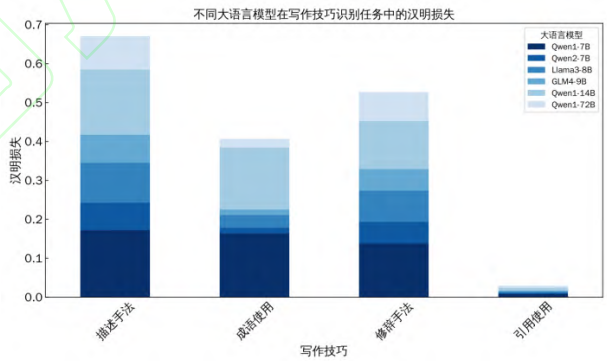

图 4 大语言模型在写作技巧维度上的汉明损失结果

Fig.4 The Hamming loss results of LLMs for writing techniques detection

基于此，本研究进一步分析各大语言模型在描写手法和修辞手法二级标签上的表现。图 5 显示了在修辞手法中，各个模型在反问和拟人方面识别效果较差。图 6 显示了各个模型在动作描写、心理描写、视觉描写、语言描写均达到良好效果，但在环境描写、外貌描写、神态描写方面整体识别水平较低。根据对写作文本的观察，环境描写、外貌描写、神态描写等技能使用的数量较少，训练数据类别不均，模型在这方面的泛化能力容易受到限制，是造成识别效率低下的原因之一。针对以上问题，模型训练可以通过持续扩展数据集写作技巧的覆盖范围，或者在训练时采用数据增强技术以增加现有数据的变化形式，以提升大语言模型对分布较少数据的泛化能力。

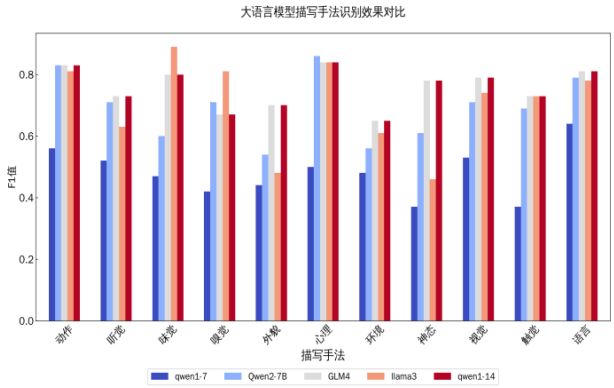

图 5 基于大语言模型的修辞识别效果

Fig.5 The effect of rhetorical recognition with LLMs

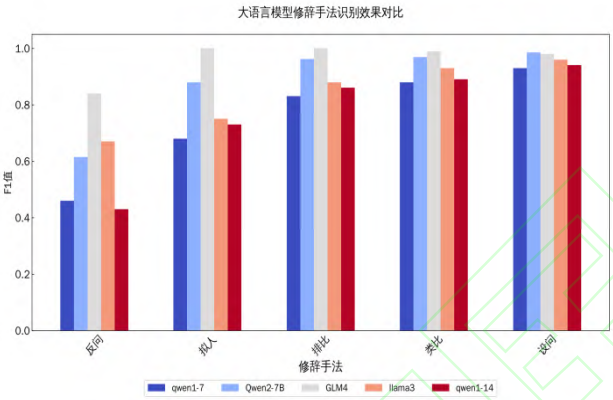

图 6 基于大语言模型的描写手法识别效果

Fig.6 The effect of description recognition with LLMs

5 结束语

本研究通过聚焦于特定写作技巧的细化分析，并结合思维链指令微调策略，为大语言模型在个性化教学领域的应用开辟了新的优化路径。具体而言，我们本地部署了 Qwen、GLM、Llama 等一系列大语言模型，并采用对话思维链指令微调方法，为学生提供精准的个性化写作反馈。实验结果表明，经过微调的模型在评语生成质量上显著优于传统分阶段训练和单轮指令微调方法。此外，本研究进一步将机器生成的评语与人类教师的评语进行了对比分析，发现机器生成的评语在情感度、信息量和连贯性方面已接近人类教师的水平，但在正确性方面仍存在提升空间。本研究提出的方法可作为实践范例，进一步融合语体、学龄、语法、语篇结构等差异化个体数据，构建更具针对性的写作智能评测指令微调数据集，将有助于更全面地满足语言教学的多样化需求，显著提升生成式人工智能在不同教学场景中的适应性和泛化能力。

参考文献

[1] MEYER J, JANSEN T, SCHILLER R, et al. Using LLMs to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions[J]. Computers and Education: Artificial Intelligence, 2024, 6: 100199.

[2] LINK S, MEHRZAD M, RAHIMI M. Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement[J]. Computer Assisted Language Learning, 2022, 35(4): 605-634.

[3] 祝智庭,张博,戴岭.数智赋能智慧教育的变与不变之道[J].中国教育信息化,2024,30(03):3-14.
ZHU Z T, ZHANG B, DAI L. The Way of Change and Constancy in Digital Intelligence Empowering Smart Education [J]. Chinese Jounal of ICT in Education, 2024, 30(03): 3-14.

[4] TAORI R, GULRAJANI I, ZHANG T, DUBOIS Y, LI X, GUESTRIN C, LIANG P, HASHIMOTO T B. Stanford Alpaca: An Instruction-following LLaMA Model[EB/OL]. 2023. https://github.com/tatsu-lab/stanford_alpaca.

[5] 张钦彤,王昱超,王鹤羲,等.大语言模型微调技术的研究综述[J].计算机工程与应用,2024,60(17):17-33.
ZHANG Q T, WANG Y C, WANG H X, et al. Com- prehensive Review of Large Language Model Fine- Tuning[J]. Computer Engineering and Applications, 2024, 60(17): 17-33.

[6] RIDLEY R, et al. Automated Cross-prompt Scoring of Essay Traits[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15).

[7] LIU Y C, HAN J W, Sboev A, et al. GEEF: A neural network model for automatic essay feedback generation by integrating writing skills assessment[J]. Expert Systems with Applications, 2024, 245: 123043.

[8] LU C, CUTUMISU M. Integrating Deep Learning into an Automated Feedback Generation System for Automated Essay Scoring[J]. International Educational Data Mining Society, 2021.

[9] MIZUMOTO A, EGUCHI M. Exploring the potential of using an AI language model for automated essay scoring[J]. Research Methods in Applied Linguistics, 2023, 2(2): 100050.

[10] YANCEY K P, LAFLAIR G, VERARDI A, et al. Rating short l2 essays on the cefr scale with gpt-4[C]// Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). 2023: 576-584.

[11] 薛嗣媛,周建设.大语言模型在汉语写作智能评估中的应用研究[J].昆明学院学报,2024,46(02):10-22.
XUE S Y, ZHOU J S. Research on the Application of Large Language Models in Intelligent Evaluation of Chinese Writing [J]. Journal of Kunming University, 2024, 46(02): 10-22.

[12] NAISMITH B, MULCAIRE P, BURSTEIN J. Automated evaluation of written discourse coherence using GPT-4[C]// Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). 2023: 394-403.

[13] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways[J]. Journal of Machine Learning Research, 2023, 24(240): 1-113.

[14] XU C W, GUO D Y, DUAN N, MCAULEY J. Baize: An open-source chat model with parameter-efficient tuning on self-chat data[J]. arXiv preprint arXiv:2304.01196, 2023.

[15] CHIANG W L, LI Z H, LIN Z, et al. Vicuna: An Open-

Source Chatbot Impressing GPT-4 with 90%*[J]. ChatGPT Quality, 2023.
[16] WEI J, WANG X Z, 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.
[17] ZELIKMAN E, WU Y H, MU J, et al. Star: Bootstrapping reasoning with reasoning[J]. Advances in Neural Information Processing Systems, 2022, 35: 15476-15488.
[18] FU Y, PENG H, OU L, SABHARWAL A, et al. Specializing smaller language models towards multi-step reasoning[C]//Proceedings of the 40th International Conference on Machine Learning, 2023:10421-10430.
[19] Mukherjee S, Mitra A, Jawahar G, et al. Orca: Progressive learning from complex explanation traces of gpt-4[J]. arXiv preprint arXiv:2306.02707, 2023.
[20] ZHANG Z X, GUAN J, XU G W, et al. Automatic comment generation for Chinese student narrative essays[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2022: 214-223.
[21] ZHENG Y W, ZHANG R C, ZHANG J H, et al. Llamafactory: Unified efficient fine-tuning of 100+ language models[J]. arXiv preprint arXiv:2403.13372, 2024.
[22] GUAN J, FENG Z E, CHEN Y, et al. LOT: A story- centric benchmark for evaluating Chinese long text understanding and generation[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 434-451.
[23] BAI J Z, BAI S, CHU Y F, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.
[24] YANG A, YANG B S, HUI B Y, et al. Qwen2 technical report[J]. arXiv preprint arXiv:2407.10671, 2024.
[25] GLM T, ZENG A H, XU B, WANG B W, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools[J]. arXiv preprint arXiv:2406.12793, 2024.
[26] WANG S Z, ZHENG Y W. Llama3-8b-chinese-chat (revision 6622a23), [CP]. (2024-04-01) [2024-11.25]. https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat.
[27] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.
[28] LIN C. Rouge: A package for automatic evaluation of summaries[C]//Text summarization branches out. 2004: 74-81.
[29] LI J W, Galley M, Brockett C, et al. A diversity- promoting objective function for neural conversation models[J]. arXiv preprint arXiv:1510.03055, 2015.
[30] ZHANG T Y Kishore V, Wu F, et al. Bertscore: Evaluating text generation with bert[J]. arXiv preprint arXiv:1904.09675, 2019.
[31] SUN Y C, LIU C, ZHOU K, et al. Parrot: Enhancing Multi-Turn Instruction Following for Large Language Models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 9729-9750.
[32] CHIANG C H, Lee H. Can large language models be an alternative to human evaluations?[J]. arXiv preprint arXiv: 2305.01937, 2023.
[33] Fleiss J L, Levin B, Paik M C. The measurement of inter-rater agreement[J]. Statistical methods for rates and proportions, 1981, 2(212-236): 22-23.
[34] LV K, YANG Y Q, LIU T X, et al. Full parameter fine-tuning for large language models with limited resources. arXiv preprint arXiv:2306.09782, 2023.
[35] HU J, SHEN Y L, WALLIS P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.