CCL24-Eval 任务8系统报告：基于指令微调与数据增强的儿童故事常识推理与寓意理解研究

于博涵，李云龙，刘涛，郑傲泽，张坤丽，咎红英
郑州大学/ 河南省郑州市
{alexyu010120, lylyun, taoliu01, zhengaoze}@gs.zzu.edu.cn
{ieklzhang, iehyzan}@zzu.edu.cn

摘要

尽管现有语言模型在自然语言处理任务上表现出色，但在深层次语义理解和常识推理方面仍有提升空间。本研究通过测试模型在儿童故事常识推理与寓意理解数据集（CRMUS）上的性能，探究如何增强模型在复杂任务中的能力。在本次任务的赛道二中，本研究使用多个7B以内的开源大模型（如Qwen、InternLM等）进行零样本推理，并选择表现最优的模型基于LoRA进行指令微调来提高其表现。除此之外，本研究还对数据集进行了分析与增强。研究结果显示，通过设计有效的指令格式和调整LoRA微调参数，模型在常识推理和寓意理解上的准确率显著提高。最终在本次任务的赛道二中取得第一名的成绩，该任务的评价指标Acc值为74.38，达到了较为先进的水准。

关键词： 儿童故事问答 ； 大语言模型 ； 指令微调 ； 数据增强

System Report for CCL24-Eval Task 8:Research on Commonsense Reasoning and Moral Understanding in Children's Stories Based on Instruction Fine-Tuning and Data Augmentation

Bohan Yu, Yunlong Li, Tao Liu, Aoze Zheng, Kunli Zhang, Hongying Zan
Zhengzhou University / Zhengzhou City, Henan Province
{alexyu010120, lylyun, taoliu01, zhengaoze}@gs.zzu.edu.cn
{ieklzhang, iehyzan}@zzu.edu.cn

Abstract

Despite the impressive performance of existing language models in natural language processing tasks, there remains significant potential for improvement in deep semantic understanding and commonsense reasoning. This study investigates methods to enhance model capabilities in complex tasks by evaluating their performance on the Children's Story Commonsense Reasoning and Moral Understanding Dataset (CRMUS). For Track 2 of this task, we employed several open-source models with fewer than 7 billion parameters (e.g., Qwen, InternLM) for zero-shot reasoning, and selected the best-performing model for instruction fine-tuning using LoRA to enhance its performance. Additionally, we conducted a thorough analysis and enhancement of the dataset. Our findings demonstrate that designing effective instruction formats and adjusting LoRA fine-tuning parameters significantly improves the accuracy of models in commonsense reasoning and moral understanding. Consequently, we achieved first place in Track 2, with an evaluation metric (Acc) score of 74.38, representing a notable advancement.

Keywords: Children's story question answering , Large language model , Instruction Tuning , Data augmentation

第二十三届中国计算语言学大会论文集，第320页-第326页，太原，中国，2024年7月25日至28日。
卷3：评测报告

1 引言

当前，自然语言处理领域对儿童故事问答这一新兴任务展现出浓厚兴趣。此任务旨在深化对儿童故事的理解与推理，为教育领域提供高效工具，促进学生理解力与语言表达技能的评估与提升。核心挑战在于深入解析给定故事与问题，检验模型对故事情节与常识的整合理解能力。具体而言，常识推理部分需依据故事情节与隐含常识，从多个选项中甄选最佳答案；寓意理解则聚焦于捕捉故事寓意，选出最贴合情节的选项。

在第二赛道中，本研究运用多种开源大型语言模型，通过开发集上的多轮零样本测试，确定InternLM2(Cai et al., 2024)为指令微调的基础模型。首先，采用固定微调策略，选取最优LoRA(Hu et al., 2021)配置。确定最佳LoRA参数后，通过不同微调模块的组合优化效果。鉴于常识推理数据类型的不平衡分布，本研究实施数据增强策略。初步利用ChatGPT生成超过200条常识推理示例，经人工审查精选137条高质量数据。将这些数据纳入开发集进行微调，显著提升了Acc指标。针对常识推理与寓意理解，各选取一组最佳参数组合，分别实现CR与MU Acc指标72.87与75.38，综合Acc指标达74.38，在本次评测中取得第一名。

2 方法

2.1 指令微调

指令微调有效强化了模型在特定任务上的表现。本研究通过精准提取数据集关键信息，构建提示模板并与数据融合，使模型深化学习儿童故事领域的专业知识，进而提升对儿童故事的语义理解与特征辨识能力。此外，经指令微调后的模型能更准确地遵循指定格式作答，极大地简化了从模型输出中抽取答案的过程。

基于零样本测试的结果，本研究在指令微调环节选用了两款以中文为主的大型语言模型——Qwen1.5-Chat-7B(Bai et al., 2023)与InternLM2-Chat-7B。其中，InternLM2-Chat-7B担任主微调模型，而Qwen1.5-Chat-7B则用于辅助验证最优LoRA参数。通过实验不同LoRA参数与微调模块搭配，最终锁定两组配置，分别对应常识推理（CR）与寓意理解（MU）的最佳Acc指标。

2.2 LoRA

由于参数规模巨大，微调整个大语言模型需要很高的成本。当用于特定任务的训练时，参数高效微调方法只需要微调少量关键参数，就能达到甚至超过全参微调的性能。其中具有代表性的是低秩适配（Low-Rank Adaptation，LoRA）方法，在冻结预训练模型权重的基础上，独立训练一个低秩分解矩阵，然后与预训练模型权重合并，方法如图1所示。

将预训练权重矩阵记为$W_0 \in R^{d\times d}$，低秩矩阵记为$\Delta W = BA$，其中$B \in R^{d\times r}$，$B \in R^{r\times d}$，d和r分别是预训练权重矩阵和低秩矩阵的秩，并且$r \ll d$，矩阵A和B分别通过随机高斯分布和零初始化，包含了可训练的参数（通常来自注意力层）。在推理阶段，使用两部分矩阵融合后的参数，如公式1所示。

$$h = W_0 x + \Delta W x \tag{1}$$

r作为一个超参数，代表了可训练参数量的规模，具体大小需要根据训练数据集大小和特点确定，此外，ΔW进一步通过超参数α缩放，决定低秩矩阵参数影响推理的程度。这种方法极大减少了内存需求，并且训练出的参数具有很强的表征能力，更适合用于特定任务的微调。

3 实验设置

3.1 数据集介绍

本研究所用数据集源于“CCL 2024 Task8儿童故事常识推理与寓意理解评测”。常识推理子任务的问题及答案经人工精心标注，而寓意理解子任务则结合自动构建与人工校验完成。常识推理涵盖社会、生物、时间、空间及物理等多元常识类型，部分题目甚至融合多种类型。数据集细分为开发集与测试集，前者含652条记录，包括400条常识推理与252条寓意理解数据；后者则拥有2768条，分别为1692条常识推理和1056条寓意理解实例。

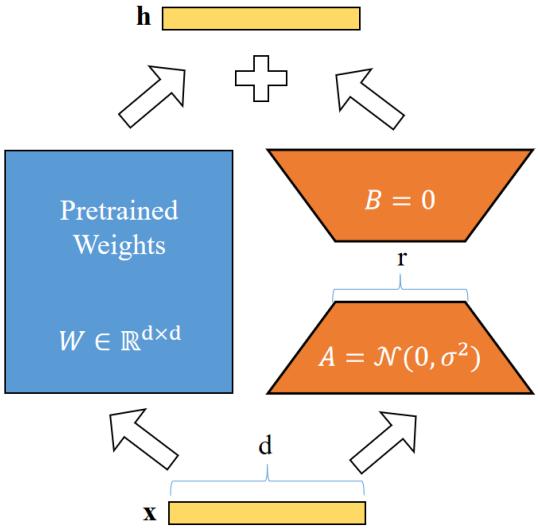

Figure 1: LoRA参数高效微调原理图

具体而言，常识推理子任务的开发集与测试集每条数据包含：标识符（id）、标题（title）、故事内容（story）、问题描述（question）、候选选项（options）、正确答案（answer）及常识类别（type）。值得注意的是，测试集的“answer”字段为空。寓意理解子任务的开发集与测试集未标注常识类型，其余属性与常识推理子任务一致。

3.2 评价指标

对于常识推理与寓意理解两项子任务，统一采用Acc指标衡量性能。参赛模型的最终评价得分由下式计算得出，即所有相关指标的加权平均：

$$Score = 0.4 \times Acc_1 + 0.6 \times Acc_2 \tag{2}$$

此处，Acc_1代表常识推理子任务的回答准确度，而Acc_2则对应寓意理解子任务的准确率。

提示模板样例

系统提示：
请根据\"生物常识、物理常识\"和一个儿童故事来做一道常识推理单项选择题。\n请你一步一步思考并直接给出答案。你将从A, B, C, D中选出正确的答案，并写在【答案】和<eoa>之间。\n完整的题目回答的格式如下：\n【答案】 ... <eoa>\n请你严格按照上述格式作答。\n儿童故事和题目如下：

用户输入：
故事:一只公鸡在田野里为自己和母鸡们寻找食物。他发现了一块宝玉，便对宝玉说：\"若不是我，而是你的主人找到了你，他会非常珍惜地把你捡起来；但我发现了你却毫无用处。我与其得到世界上一切宝玉，倒不如得到一颗麦子好。\"\n题目:关于公鸡对宝玉的看法，下列选项描述正确的是？\nA. 宝玉太硬了，不好吃 B. 主人非常喜欢吃宝玉 C. 宝玉不是食物，但自己可以拿去卖钱 D. 宝玉不是食物，不能吃

期望输出：
【答案】 D <eoa>

Figure 2: 指令数据示例

3.3 数据预处理

鉴于原始数据集包含若干非必要字段，如id、title、domain等，本研究对寓意理解（MU）数据集进行了精简，仅保留story、question、options与answer字段。相比之下，常识推理（CR）任务因涉及特定类型常识，故在保留前述字段的基础上，额外保存type字段，确保模型能依据常识类别提供更为精确的回应。

提示模板的设计借鉴了GAOKAO-Bench(Zhang et al., 2023)的研究，最终定制的指令模板详情见图2。

3.4 数据增强

鉴于开发集内常识推理任务各类常识分布的不均衡性，本研究引入数据增强技术予以应对。表1展示了各常识类别的样本量，显而易见，空间与物理常识远少于社会及生物常识。模型在开发集上的推理结果显示，其在时间、空间与物理常识的表现欠佳，推测原因可能与这些类型数据稀缺有关。为此，本研究首先借助ChatGPT生成逾200条常识推理示例，再经人工严格筛选，最终精选137条优质数据。增强后的各类常识数量详情参见表2。

常识类型	社会常识	生物常识	时间常识	空间常识	物理常识
常识数量	196	90	67	37	37

Table 1: 开发集上各常识类型的数量

常识类型	社会常识	生物常识	时间常识	空间常识	物理常识
常识数量	212	138	110	51	67

Table 2: 扩充开发集后各常识类型的数量

4 实验流程

4.1 实验参数设置

本研究的实验参数配置详列于表3。所有实验采用PyTorch深度学习框架执行，微调工作依托于LLaMA Factory(Zheng et al., 2024)框架，并在配备一张4090与一张A40的硬件环境下运行。

模型参数	参数值
训练轮数	5
学习率	5e-5
截断长度	1536
Batch size	1
Optimizer	AdamW
Warmup ratio	0.1
Lr scheduler	Cosine
Gradient accumulation steps	8

Table 3: 实验参数

4.2 模型选择

为确保在性能优越的模型基础上开展微调，本研究挑选了Baichuan2-Chat-7B(Yang et al., 2023)、Qwen-Chat-7B、Qwen1.5-Chat-7B、Yi-Chat-6B(Young et al., 2024)与InternLM2-Chat-7B，在原始开发集（含400条常识推理[CR]数据与252条寓意理解[MU]数据）上执行零样

本推理。表4呈现的实验成果表明，所有测试模型在CR与MU任务上的Acc指标均超越主办方提供的基准线。综合考量各模型的Acc得分后，最终选定InternLM2-Chat-7B作为微调工作的基准模型。

模型	CR	MU	Overall
Baichuan2-Chat-7B	43.5	42.8	43.1
Qwen-Chat-7B	45.5	40.4	42.4
Qwen1.5-Chat-7B	62.5	38.0	47.8
Yi-Chat-6B	50.3	44.4	46.8
InternLM2-Chat-7B	65.7	52.7	57.9
Baseline	31.2	33.2	32.4

Table 4: 各模型在开发集上零样本推理的表现

4.3 LoRA参数选择

LoRA的超参数设定对实验成效具有显著影响。本研究参照LoRA论文建议，于原始开发集上实施多组超参数微调，并在测试集上评估结果，详情见表5。其中，Wqkv构成基本微调模块。实验显示，当Rank设为64、Alpha为16时，模型表现欠佳；调整至LoRA论文推荐的Rank：Alpha比例1：2后，模型性能显著提升。尤其当Rank等于256、Alpha设定为512时，模型效能达到顶峰。

除在InternLM-Chat-7B上探索最优参数外，本研究亦采用Qwen1.5-Chat-7B在原始开发集上执行多轮微调试验，测试集上的结果列于表6。综合考量InternLM-Chat-7B与Qwen1.5-Chat-7B的实测表现，最终决定采用Rank=256、Alpha=512的配置。

Rank	Alpha	Target	CR	MU	Overall
64	16	Wqkv	65.90	52.75	58.01
256	512	Wqkv	70.39	71.21	70.88
512	512	Wqkv	69.73	71.40	70.73
512	1024	Wqkv	69.56	70.45	70.09

Table 5: InternLM-Chat-7B的不同LoRA超参数设置

Rank	Alpha	Target	CR	MU	Overall
32	64	Wqkv	34.86	67.04	54.17
64	128	Wqkv	64.42	68.46	66.84
128	256	Wqkv	65.95	69.31	67.97
256	512	Wqkv	65.36	69.6	67.9

Table 6: Qwen1.5-Chat-7B的不同LoRA超参数设置

4.4 LoRA微调模块选择

依据4.3章节中选定的最优LoRA超参数，本研究进一步探索不同模块组合的效果。InternLM2-Chat-7B模型可供微调的组件涵盖Wqkv、W1、W2、W3、Wo，其中Wqkv被视为基础微调模块，其余则为可选附加模块。将Wqkv与任一额外模块搭配进行微调实验，所得测试集上的结果详载于表7。

审视表7可发现，Wqkv与W1的组合展现出最优效用。鉴于模块叠加可增扩微调参数规模，从而增强模型效能，本研究特增设一组实验，测试Wqkv联合W1、W2的效果。相应结果收录于表8。

Rank	Alpha	Target	CR	MU	Overall
256	512	Wqkv	70.39	71.21	70.88
256	512	Wqkv,W1	72.1	71.78	71.91
256	512	Wqkv,W2	71.21	72.15	71.77
256	512	Wqkv,W3	70.68	72.34	71.68
256	512	Wqkv,Wo	70.98	72.15	71.68

Table 7: 在原始开发集上加入单一模块的微调

Rank	Alpha	Target	CR	MU	Overall
256	512	Wqkv,W1,W2	72.87	71.78	72.22

Table 8: 在原始开发集上加入W1、W2模块

模型在整合W2后，展现出性能的显著提升，这一进展激励了对更多模块组合的探索。本研究在扩充后的数据集上进行了多轮精细的微调实验。根据表9中的实验结果，数据增强虽使CR任务的表现略有下降，却提升了MU任务的性能。分析认为，CR任务涉及的推理过程较为繁复，现有增强数据的复杂程度可能不足以完全匹配其需求；相反，MU任务侧重于直接从文本中提取答案，因此，增强的CR信息有效地深化了MU情境下文本的信息层次。

Rank	Alpha	Target	CR	MU	Overall
256	512	Wqkv,W1,W2	69.9	74.14	72.44
256	512	Wqkv,W2,W3	70.8	74.9	73.26
256	512	Wqkv,W1,W3	70.56	75.38	73.45
256	512	Wqkv,W2,Wo	71.57	72.91	72.37
256	512	Wqkv,W1,W2,W3	71.04	72.25	71.76
256	512	Wqkv,W1,W2,Wo	71.74	72.15	71.99
256	512	Wqkv,W2,W3,Wo	67.96	71.02	69.8
256	512	Wqkv,W1,W2,W3,Wo	71.74	71.59	71.65

Table 9: 在扩充后的开发集上微调更多模块

通过对不同模块进行实验对比，观察到模型性能并未随微调参数数量的增加而持续上升。令人意外的是，当所有可微调模块均被纳入时，模型表现出现下滑。基于此，本研究最终采纳了表8中记录的最优CR结果，以及表9中所列的最佳MU结果，具体详情参见表10。

CR	MU	Overall
72.87	75.38	74.38

Table 10: 最佳分数组合

5 结论

本研究深入分析并评估了语言模型在处理复杂语义理解，特别是常识推理与寓意理解任务时的能力提升途径。在第二赛道的竞赛中，我们选择了Qwen1.5-Chat-7B与InternLM2-Chat-7B等模型作为研究对象。通过参数优化与数据增强技术的应用，发现InternLM2-Chat-7B在上述两项任务中展现出最优性能。实验成果证实，精心设计的提示模板结合LoRA微调参数调整及数据增强策略，能显著提高模型的常识推理精度和寓意理解水平，有力地促进了语言模型在

复杂语义理解和常识推理领域的发展。这一研究为后续在自然语言理解领域，尤其是儿童故事分析方向，奠定了坚实的基础，提供了宝贵的指导思路。

参考文献

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.