

文章编号:1671-5896(2026)04-0863-09

基于思维链的问答系统增强推理方法

曹茂俊, 王亚飞, 肖红, 彭超, 杨家晨

(东北石油大学 计算机与信息技术学院, 黑龙江 大庆 163318)

摘要: 针对生产测井数据多源异构、解释多解性强和传统方法依赖专家经验且效率较低等问题, 提出了一种基于思维链(CoT: Chain of Thought)的问答系统增强推理方法。该方法在大语言模型元 AI(LLaMA2: Large Language Model Meta AI2)基础上, 引入参数高效微调低秩适配(LoRA: Low-Rank Adaptation)技术, 通过在自注意力和前馈网络层插入低秩矩阵, 实现了模型性能保持与计算资源优化的平衡。同时, 构建了发散思维链(DCoT: Divergent Chain of Thought)推理框架, 使模型能生成多条推理路径, 提升系统对复杂生产测井问题的多步解释能力。实验表明, 该方法在 F_1 、精确率和召回率上分别提升 4.6%、7.0% 和 2.1%, 验证了其推理性能的优越性。研究证明, LoRA+DCoT 方法在提升测井数据智能解释方面具有良好的实用性和可扩展性。

关键词: 大语言模型; 思维链; 生产测井; 增强推理; 低秩适配; 发散思维链

中图分类号: TP391

文献标志码: A

Chain-of-Thought Based Enhancement Method for Inference in Question-Answering Systems

CAO Maojun, WANG Yafei, XIAO Hong, PENG Chao, YANG Jiachen

(School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: To address the issues of multi-source heterogeneity in production logging data, strong multi-solution interpretability, the inefficiency and expert-dependency of traditional methods, a question-answering system enhanced reasoning method based on CoT(Chain of Thought) is proposed. This method is built on the LLaMA2 large language model and the parameter-efficient fine-tuning LoRA (Low-Rank Adaptation) technique is introduced, which inserts low-rank matrices into the self-attention and feed-forward network layers to achieve a balance between model performance maintenance and computational resource optimization. A DCoT(Divergent Chain of Thought) reasoning framework is constructed, enabling the model to generate multiple reasoning paths and enhancing the system's multi-step interpretation ability for complex production logging problems. Experimental results show that this method improves F_1 score, precision, and recall by 4.6%, 7.0%, and 2.1%, respectively, compared to the baseline model, demonstrating strong reasoning performance and application potential. The research proves that the LoRA+DCoT method has good practicality and scalability in improving the intelligent interpretation effect of logging data.

Key words: large language model; chain of thought; production logging; enhanced reasoning; low-rank adaptation(LoRA); divergent chain of thought(DCoT)

收稿日期: 2025-07-21

基金项目: 国家自然科学基金资助项目(42172161; 52474035); 黑龙江省自然科学基金联合基金重点资助项目(ZL2024D003); 中石油创新基金资助项目(2024DQ02-0114)

作者简介: 曹茂俊(1978—), 男, 黑龙江大庆人, 东北石油大学副教授, 主要从事深度学习、智能计算在测井曲线识别中应用研究, (Tel)86-13796988520(E-mail)caomaojun@126.com; 通讯作者: 王亚飞(1999—), 男, 河南安阳人, 东北石油大学硕士研究生, 主要从事深度学习、人工智能研究, (Tel)86-18639313868(E-mail)wyf339966@163.com。

0 引言

随着 Transformer^[1]架构的提出与优化,大规模预训练语言模型(LLMs: Large Language Models)在自然语言处理领域取得了显著进展,其不仅能生成高质量的文本,还展现出强大的推理与理解能力。而如何进一步提升大模型的推理深度与可解释性,成为当前人工智能研究的热点问题。在传统重工业领域如石油勘探与开发中,尤其是生产测井解释环节,人工智能^[2]技术的应用仍处于探索阶段。当前,生产测井数据通常表现为多源异构性与强多解性,其解释过程高度依赖专家经验,存在解释周期长、效率低、主观性强等问题,难以满足智能化油田建设的迫切需求。虽然已有部分研究尝试引入深度学习^[3]方法辅助解释,但在多模态信息融合、推理链构建以及系统交互性等方面仍存在显著不足。

CoT(Chain of Thought)作为一种新兴的推理范式,通过引导模型显式生成中间推理步骤,显著增强了其在多步逻辑任务中的表现力与稳定性。Kojima 等^[4]提出了零样本思维链(ZSCoT: Zero-Shot CoT),通过引入“让我们一步一步地思考”的提示,显著提升了模型在推理任务中的表现,展示了预训练语言模型在无任务示例情况下的推理潜力。2022年,Wei 等^[5]研究了语言模型的涌现能力,发现这些能力在模型达到一定规模后才显著显现,并探讨了进一步扩展模型规模可能带来的新能力。2023年,Wang 等^[6]提出了计划和解决(PS: Plan-and-Solve)提示策略,通过将任务划分为子任务并逐步执行,解决了 ZSCoT 在多步推理中的遗漏步骤问题。2024年,Yue 等^[7]提出利用广泛认可的多模态推理基准(MMMU: Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark),并在 MMMU 上进行多模态 CoT 评估,无需进一步培训。但上述研究都是基于单条思维链的推理模式,虽然提升了模型的逻辑连贯性,但在处理开放域问题或需要创造性解决方案的任务时仍存在局限性。与传统 CoT 的单一路径推理不同,DCoT(Divergent Chain of Thought)通过生成多条推理路径,模拟人类专家的发散性思维,显著提升了模型对多解性问题的覆盖能力,尤其适用于测井数据解释中的复杂场景。

为此,笔者提出一种基于思维链的问答系统增强推理方法,结合 LLaMA2(Large Language Model Meta AI2)等先进预训练语言模型,引入 LoRA(Low-Rank Adaptation)与 DCoT 推理机制,实现对测井问题的多路径、多步骤智能解释。其不仅提升了模型在复杂工业任务中的适应性与推理精度,也为构建高效、可交互、具备可解释性的智能生产测井系统提供了新思路与关键技术支持。

1 预备知识

1.1 测井基础知识

生产测井是油气田开发和动态监测中的核心技术,主要用于评估井筒内流体性质、产能分布、井筒完整性及其变化趋势。常见的生产测井的数据以曲线、表格和文档等形式存在,具有连续性、时效性强、多源异构等特点。

测井解释是对测井数据进行定量和定性分析的过程,传统方法多依赖人工经验判断,存在效率低、主观性强的问题。随着数据体量增大和解释任务复杂化,传统软件系统逐渐难以满足现场对高效、智能和可解释系统的需求。因此,融合自然语言处理、知识图谱^[8]与深度学习^[9]等先进技术,提升测井解释的智能化水平。

1.2 LLaMA2 语言模型

大语言模型元 AI(LLaMA)是 Meta 公司推出的一系列开源大语言模型^[10],旨在推动自然语言处理领域的研究与应用。其最新版本 LLaMA2 在架构设计和训练策略上进行了多项优化,显著提升了模型的性能和泛化能力。LLaMA2 基于 Transformer 解码器架构,融合了预归一化、旋转位置编码以及高效注意力机制^[9]等先进技术,使其在中等参数规模下实现了卓越的表现和更稳定的训练过程。

在实际应用中,开源模型 LLaMA2^[11]展现出强大的文本生成、理解和推理能力,广泛适用于包括问答系统^[12]、对话生成、文本摘要、信息抽取等多种下游任务。LLaMA2 通过其优异的泛化能力和强大的

上下文推理能力,可以有效提升测井解释的准确性和自动化水平,促进智能化油气勘探与开发的进步。此外,LLaMA2 较为轻量的模型规模和高效的计算特性,也使其更易于在实际生产环境中部署和扩展,降低了硬件和能耗成本。

1.3 模型微调与 LoRA 技术

预训练语言模型虽然具备较强的通用能力,但在特定领域任务中通常表现有限。微调技术通过在目标任务上调整模型参数,显著提升专业场景下的表现。目前微调方法主要分为全参数微调(Full Fine-tuning)和参数高效微调(PEFT: Parameter-Efficient Fine-Tuning),并且后者能以更低的训练成本实现接近全参微调的效果。

LoRA^[13]作为典型的 PEFT 技术,其核心思想是在冻结预训练模型参数的基础上,于 Transformer 的注意力层和前馈层插入低秩可训练矩阵,实现参数高效更新。LoRA 不仅大幅减少显存消耗和训练时间,还能在推理阶段将增量权重与原模型权重合并,并不影响推理效率和部署复杂度。因此,笔者选用LoRA 技术作为模型微调的主要方案,以提升测井问答系统的推理效率与资源适配能力。图 1 给出了 LoRA 的模型结构,通过在注意力层插入低秩矩阵(A、B)实现参数高效更新。图 2 为 LoRA 微调原理示意图,显示增量权重与原权重的合并过程。

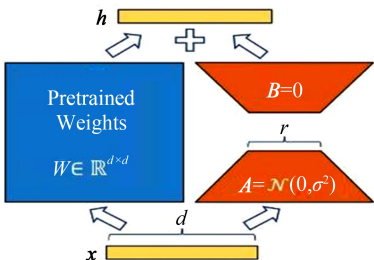


图 1 LoRA 模型结构

Fig. 1 LoRA model structure diagram

LoRA 通过向 Transformer 层注入两个低秩矩阵 A 和 B,将传统全参数微调中庞大的参数量压缩数个数量级。微调过程中,预训练区中 W_0 固定不变,增量权重通过低秩矩阵计算:

$$h = (W_0 + \Delta W)x = W_0x + BAx, \tag{1}$$

其中 x 为输入向量, h 为输出向量, W_0 为冻结的预训练权重, $A \in \mathbb{R}^{r \times d}$ 、 $B \in \mathbb{R}^{d \times r}$ 为可训练的低秩矩阵, $r \ll \min(d, k)$, α 为缩放因子, d 为输出维度(W_0 的行数), k 为输入维度(W_0 的列数)。训练时,仅更新 A 和 B 参数,保持 W_0 不变;推理时,将 BA 与 W_0 合并,实现零推理延时(见图 2)。

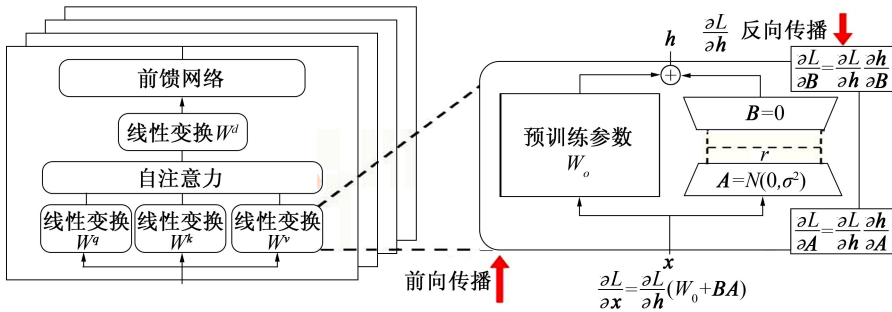


图 2 LoRA 微调原理

Fig. 2 LoRA fine tuning schematic

1.4 思维链推理

思维链推理是一种模拟人类逐步展开思考过程的方法,旨在通过引导大型语言模型生成清晰、连贯的中间推理步骤,从而提升模型在处理复杂任务时的可解释性和准确性。然而,传统的思维链通常依赖于单一路径的线性推理流程,在面临多解性强的任务时,存在明显的局限性。单一推理路径容易忽略潜在的多解法,限制了模型探索问题全貌的能力,特别是在专业领域如生产测井数据解释中,因地质条件复杂多变,问题本身通常具备多种合理解释。因此引入发散思维链解决多路径问题。具体示例如图 3 所示。

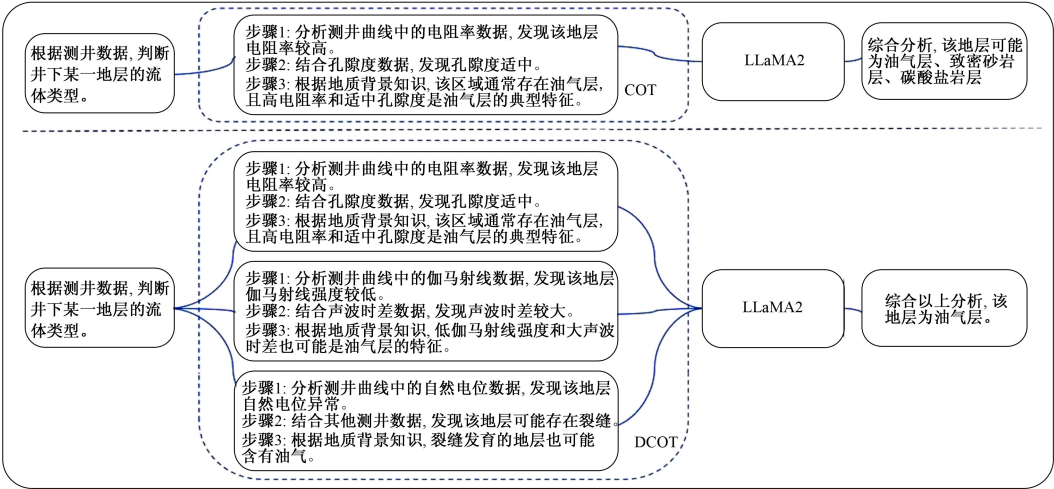


图 3 生产测井思维链应用实例

Fig. 3 Application example of chain-of-thought in production logging

2 发散思维链增强推理方法

针对生产测井数据多源异构且存在强多解性的特点,笔者提出了一种基于发散思维链^[14]的增强推理方法。该方法以大语言模型 LLaMA2 为基础,结合 LoRA 与多路径推理策略,有效实现了测井问答系统的多步逻辑推理和多样化答案生成。整体方案涵盖数据分析、LoRA 微调配置、发散思维链构建以及模型性能评估四大核心环节,旨在提升系统对复杂测井问题的推理能力和适应性。

2.1 数据分析与预处理

生产测井^[15]数据通常由曲线图、报表表格及解释文档等构成,并具有如下典型特征:

- 1) 连续性与多解性:测井曲线连续性强,但因地层复杂,同一曲线可能对应多种地质解释;
- 2) 结构化与非结构化并存:表格与图像数据结构清晰,而文档和说明内容多为非结构化文本;
- 3) 多源异构:数据来源多样,包括生产平台、地质解释系统与人工记录等,格式兼容性要求高。

为此,笔者设计数据融合机制,将曲线、表格、文本等信息统一映射到语义空间,为后续问答任务提供语义支撑。

2.2 LoRA 配置与微调方法

在大语言模型中,Transformer 是基础架构,其中注意力模块尤为关键。微调时,固定预训练参数,仅对 LoRA 参数进行训练。设预训练权重为 W_0 ,训练时不更新该权重,而是在其基础上学习两个低秩矩阵 A 和 B ,最终权重通过以下方式合并:

$$W = W_0 + \alpha BA,$$

其中 α 为缩放因子,用于控制 LoRA 对模型的影响强度, A 和 B 的秩为 r 。笔者将 LoRA 模块插入到注意力机制中的 Q 、 K 、 V 3 个线性映射层,如图 4 所示。

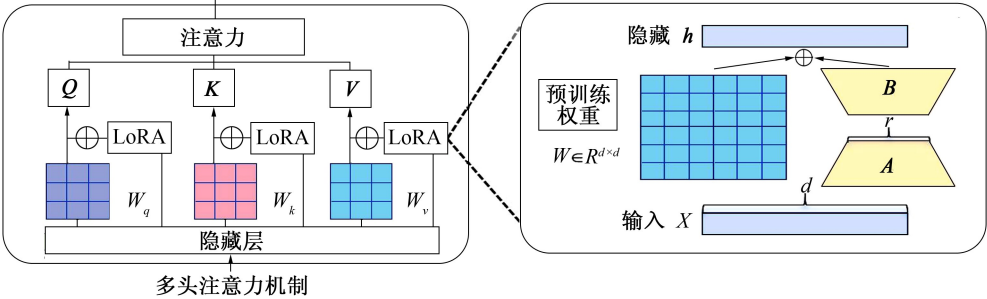


图 4 LoRA 在注意力层中的低秩适配结构

Fig. 4 LoRA's low-rank adaptation structure in the attention layer

缩放因子 α 是关键调节参数, 其决定了 LoRA“补丁”对模型行为的影响大小。为验证不同 α 取值对训练效果的影响, 笔者在数据集上进行了对比实验, 结果如表 1 所示。

表 1 不同 α/r 比例的鲁棒性分析

Tab. 1 Robustness analysis for different α/r ratios

α/r	训练稳定性(梯度方差↓)	准确性	α/r	训练稳定性(梯度方差↓)	准确性
0.5	极高	58.1	2.0	中	65.1
1.0	高	63.5	4.0	不稳定(梯度裁剪触发率↑)	62.3

当 $\alpha/r=2$ 时, 表明参数更新更平滑, 有效避免训练震荡, 并且准确性更高。对秩参数 r , 其决定了 LoRA 低秩矩阵的维度, 常见取值有 1、4、8、16、32。秩的大小直接影响微调效果与参数量, 过大并不一定带来更优表现。表 2 给出了在生产测井数据集上, 不同 LoRA 权重插入方式和秩参数的效果对比。将 LoRA 同时应用于 $Q、K、V$ 层, 且设定秩为 8, 能获得最佳的微调性能。

表 2 生产测井数据集在不同权重和不同秩的效果对比

Tab. 2 Comparison of the effects of production logging datasets at different weights and different ranks

权重类型	$r=1$	$r=4$	$r=8$	$r=16$	$r=32$
W_q	50.4	51.3	52.1	53.4	52.5
60.2	W_q, W_v	56.4	60.3	61.2	59.5
63.7	W_q, W_k, W_v	62.1	63.5	65.2	64.1

2.3 发散思维链构造方法

传统的 CoT 采用单一的线性流程, 易导致遗漏其他潜在推理路径。为提升模型在复杂问题上的推理覆盖度, 笔者引入了 DCoT 构造策略, 如图 5 所示。主要包括以下核心内容:

- 1) 提示模板设计: 在输入中明确指定生成多个思维路径的数量 k , 格式为“[问题]+[选项]+[答案数 k]”, 模型输出结构为“[答案 1: CoT1]+...+[答案 k : CoTk]+[最终答案]”;
- 2) 推理多样性训练: 通过多条不同推理路径对同一问题进行训练, 增强模型对的泛化能力;
- 3) 路径选择机制: 基于语义相似度和得分策略, 从生成的多条思维链中自动筛选最优路径。

该 DCoT 策略模拟人类专家的发散性思维方式, 显著提升问答系统对不确定性和模糊性问题的适应力, 尤为适合多解性强的测井解释任务。

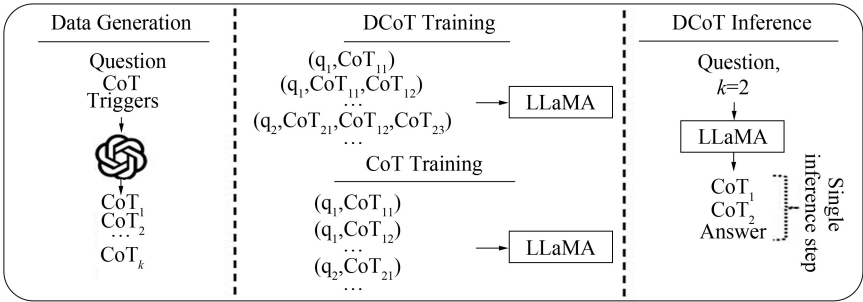


图 5 DCoT 的训练与推理流程

Fig. 5 Training and inference process for DCoT

2.4 推理效果评估指标

为全面评估基于思维链的问答系统推理性能, 笔者构建了多维度、多层次的评估体系, 涵盖准确性、鲁棒性、效率及可解释性 4 个方面。具体指标如下。

Precision(精准率)表示模型预测正确的答案占有所有预测答案的比例, 反映答案的准确度:

$$P_{\text{Precision}} = \frac{T_p}{T_p + F_p} \circ$$

Recall(召回率)表示模型预测正确答案占有所有真正正确答案的比例, 衡量模型识别正确答案的能力:

$$R_{\text{Recall}} = \frac{T_p}{T_p + F_N},$$

其中 T_p (True Positive) 为模型正确预测的答案数量, F_p (False Positive) 为模型错误预测的答案数量, F_N (False Negative) 为模型未能预测出的真实答案数量。

采用 F_1 分数作为精准率与召回率的调和平均数, 用于综合衡量答案生成的完整性与精确性。 F_1 分数的计算基于语义级别的匹配, 具体通过预训练语言模型生成问题与答案的嵌入向量, 利用余弦相似度判定语义匹配程度, 如下:

$$F_1 = 2 \times \frac{P_{\text{precision}} R_{\text{Recall}}}{P_{\text{Precision}} + R_{\text{Recall}}}。$$

3 实验对比与分析

为验证笔者提出的基于 LoRA 与 DCoT 方法在生产测井问答任务中的有效性, 设计并实施了一系列对比和消融实验。实验涵盖数据集构建、模型训练、性能对比与策略评估等环节, 以全面分析该方法在实际应用场景下的表现。

3.1 数据集构建

笔者所用数据集来源于大庆油田现场生产测井数据, 涵盖测井解释报告、生产报表、曲线数据及图件说明等多种文档类型。为提升大语言模型的推理能力, 数据统一转换为指令微调格式, 构建了包含 1 万条高质量问答的多路径思维链数据集。该数据集集成了多模态融合, 通过描述性语言将表格、文本与图像数据转化为自然语言问题; 针对多解性特点设计了多个合理思维链路径, 增强模型复杂推理覆盖能力; 且问题设计紧贴井场实际需求, 如层位产能分析、流体类型判断及完井层段解释, 具有较强的应用价值。数据集按 8 : 1 : 1 比例划分为训练、验证和测试集, 确保评估的客观性和稳定性。数据集属性示例如表 3 所示。

表 3 数据集示例
Tab.3 Data set example

序号	属性	说明	示例
1	id	唯一序号	4 587
2	question	题干	测井数据电阻率: 20 Ω · m, 孔隙度: 18%, 含水率: 45%。 根据这些数据, 判断该地层的流体类型。
3	context	上下文	地层的流体类型大致分为油层、气层、水层、油气层等。
4	answer	标准答案	油气层
5	correct_cots	多条思维链	1) 基于电阻率分析。2) 基于含水率分析。3) 综合考虑 3 个参数分析。

3.2 模型训练配置

训练模型基于开源 LLaMA2-7B 架构, 采用 PyTorch 框架和 Transformers 库实现。为降低训练成本并提升迁移能力, 微调采用 LoRA 技术, 核心配置包括: 在阿里云搭载 NVIDIA Tesla A100 40 GByte 的 Graphics Processing Unit 节点上进行训练, 使用 AdamW 优化器和线性学习率衰减调度, LoRA 插入于 Q 、 K 、 V 3 个层, 秩参数通过实验确定为 $r=8$, $\alpha/r=2$ 。训练时长约 100 h, 直至 loss 收敛, 期间模型表现稳定, 无显著过拟合或梯度爆炸。训练过程通过 Wandb 进行监控, 实时跟踪 loss、精度及学习率等指标, 验证了 LoRA 配置的有效性和适用性。

3.3 对比实验

为全面评估笔者所提出的基于思维链增强推理方法在生产测井问答任务中的效果, 选取当前主流且公开可获得的多个大规模开源模型作为对比对象。具体包括 OLMo^[16] (基于解码器-only Transformer 架构, 并在多个硬件类型上进行了训练)、Bloom^[17] (一个多语言的大型语言模型 (LLM), 拥有 1 760 亿个参数, 能以 46 种自然语言和 13 种编程语言生成文本)、MosaicML Pretrained Transformer^[18] (一种 decoder-style 的 model, 每个 query 只能访问到在它之前的 value)、Falcon^[19] (仅使用互联网语料, 不需要额外数据

源就可以训练一个不错的大模型)、Gemma^[20](GoogleDeepMind 和 Google 其他团队开发,灵感来自 Gemini,并且在发布模型权重的同时发布了支持开发者创新、促进协作以及指导负责使用 Gemma 模型的工具)、Mistral^[21](用于复杂的多语言推理任务,包括文本理解、转换和代码生成)等 6 个 7B 规模的模型,以及业界较为成熟且广泛使用的 LLaMA2-7B 模型。所有模型均在相同的硬件环境下,使用相同的测井问答数据集进行推理测试,确保测试条件公平统一,实验结果如表 4 所示。

表 4 实验结果

Tab. 4 Experimental results

Model	参数	Precision	Recall	F_1	Model	参数	Precision	Recall	F_1
OLMo-7B	7B	56.5	55.0	55.7	Gemma-7B	7B	60.8	58.9	59.8
Bloom-7B	7B	55.9	54.1	55.0	Mistral-7B	7B	61.2	59.8	60.5
MPT-7B	7B	58.0	56.2	57.1	LLaMA2-7B	7B	63.5	62.1	62.8
Falcon-7B	7B	57.1	55.3	56.2					

从表 4 可看出,虽然各模型性能相近,但笔者提出的基于 LLaMA2 结合 LoRA 微调 and 思维链推理机制的方法,显著超越其他所有模型。OLMo 模型在科研透明性上突出,但在需要领域自适应的生产场景中,其性能劣势明显;Bloom 在多语言任务上性能优异,但在以中文测井报告为主的场景下,其跨语言泛化能力反而导致领域偏差;MPT 的 8k 上下文并未带来显著增益,说明测井问答对长依赖需求相对有限,而更依赖细粒度数值-文本对齐;Falcon 与 Gemma 训练语料以通用网络文本为主,缺乏石油工程领域知识,导致在“流体类型识别”类题目的召回率普遍低于 55%。

虽然这些开源模型在各自的应用场景中可能表现出色,但在生产测井问答这种特定任务上,其并未能充分挖掘测井数据中的复杂逻辑关系和多解性特点。而笔者方法通过 DCoT 的构造策略,能有效地探索更广泛的解空间,从而更好地应对测井数据的多解性问题,生成更准确、全面的答案。

3.4 消融实验及分析

为进一步剖析本文各关键模块对整体模型性能的贡献,笔者设计并实施了 4 组消融实验,系统评估模型组合条件下对精准率、召回率及 F_1 分数的影响。消融实验具体方案如下:Original(LLaMA2 原始模型):未经任何微调和推理机制增强的基础模型,作为对照组。LLaMA2+LoRA 微调:引入 LoRA 技术对模型进行参数高效微调,以适应测井问答任务。LLaMA2+LoRA+CoT 线性思维链:在 LoRA 微调基础上,加入了 CoT,帮助模型线性组织推理步骤。LLaMA2+LoRA+DCoT:结合 LoRA 微调与 DCoT,以增强模型的推理灵活性和答案多样性。实验结果如表 5 所示。

表 5 消融实验结果

Tab. 5 Ablation results

Model	Precision	Recall	F_1	Model	Precision	Recall	F_1
Original	56.5	60.0	58.2	Original+LoRA+CoT	59.0	61.6	60.3
Original+LoRA	57.8	60.5	59.1	Original+LoRA+DCoT	63.5	62.1	62.8

从表 5 可看出,基线模型的 F_1 为 58.2%,表现为基础水平。引入 LoRA 微调后,模型在精准率和召回率均有所提升, F_1 分数提升至 59.1%。进一步将 CoT 线性思维链引入模型,模型的推理能力得到显著增强, F_1 达到 60.3%,精准率和召回率均有约 1% 的提升。结合 LoRA 与 DCoT 的技术方案实现了最优性能, F_1 分数达到 62.8%,精准率和召回率分别提升至 63.5% 和 62.1%。

通过对实验结果进行分析,可以发现 LoRA 技术在预训练模型的关键层插入低秩矩阵,显著降低了参数更新的计算成本,同时保持了模型的高性能。实验表明,仅用 0.16% 的参数量就实现了相当于全参数微调 97% 的效果,表明 LoRA 在资源优化和性能提升方面具有显著优势。DCoT 通过引导模型生成多条推理路径,有效提升了模型对复杂问题的多解性处理能力。实验中,DCoT 显著提高了模型的 F_1 分数、精准率和召回率,表明其在提升模型推理多样性和答案全面性方面发挥了重要作用。通过这些方法的有机结合,模型在生产测井问答任务中表现出了显著的优越性。

4 结 语

笔者提出的基于 LoRA 微调与 DCoT 的方法,在 LLaMA2-7B 模型上实现了生产测井问答任务的较优性能, F_1 分数达到了 62.8%。通过引入发散思维链,模型能探索更广泛的解空间,有效应对测井数据中存在的多解性问题,从而提升答案的全面性与准确性。

展望未来,计划进一步研究自适应 k 值的动态选择策略,提升发散思维链的推理灵活性和精度;并探索多模态 DCoT 方案,融合测井曲线等图像信息,拓展模型对多源异构数据的综合推理能力,希望进一步增强系统的实用性和应用广度。

参考文献:

[1] 钱亮宏,王福德,孙晓海. 基于预训练 Transformer 语言模型的源代码剽窃检测研究 [J]. 吉林大学学报(信息科学版), 2024, 42(4): 747-753.
QIAN L H, WANG F D, SUN X H. A Study on Source Code Plagiarism Detection Based on Pre-Trained Transformer Language Model [J]. Journal of Jilin University (Information Science Edition), 2024, 42(4): 747-753.

[2] 袁杰,王军博,陈歆,等. 人工智能在超高性能混凝土中的应用研究进展 [J]. 吉林大学学报(工学版), 2025, 55: 771-789.
YUAN J, WANG J B, CHEN X, et al. Research Progress on the Application of Artificial Intelligence in Ultra-High Performance Concrete [J]. Journal of Jilin University (Engineering and Technology Edition), 2025, 55: 771-789.

[3] 张汝波,常世淇,张天一. 基于深度学习的图像信息隐藏方法综述 [J/OL]. 吉林大学学报(工学版), 1-19 [2026-05-04]. <https://link.cnki.net/doi/10.13229/j.cnki.jdxbgxb.20240381>.
ZHANG R B, CHANG S Q, ZHANG T Y. A Review of Image Information Hiding Methods Based on Deep Learning [J/OL]. Journal of Jilin University (Engineering and Technology Edition), 1-19 [2026-05-04]. <https://link.cnki.net/doi/10.13229/j.cnki.jdxbgxb.20240381>.

[4] KOJIMA T, GU S S, REID M, et al. Large Language Models Are Zero-Shot Reasoners [J]. Advances in Neural Information Processing Systems, 2022, 35: 22199-22213.

[5] WEI J, TAY Y, BOMMASANI R, et al. Emergent Abilities of Large Language Models [J/OL]. 2022 [2026-05-04]. <https://arxiv.org/abs/2206.07682>.

[6] WANG L, XU W, LAN Y, et al. Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models [J/OL]. 2023 [2026-05-04]. <https://arxiv.org/abs/2305.04091>.

[7] YUE X, NI Y, ZHANG K, et al. MMMU: A Massive Multi-Discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024: 9556-9567.

[8] 陈正非,席晶,李志勇,等. 基于多源数据融合的电建知识图谱构建研究 [J]. 吉林大学学报(信息科学版), 2024, 42(5): 921-929.
CHEN Z F, XI S, LI Z Y, et al. Research on the Construction of Electric Construction Knowledge Graph Based on Multi-Source Data Fusion [J]. Journal of Jilin University (Information Science Edition), 2024, 42(5): 921-929.

[9] 彭建祥. 深度学习模式下大数据特征集成分类算法 [J]. 吉林大学学报(信息科学版), 2025, 43(2): 231-237.
PENG J X. A Feature-Integrated Classification Algorithm for Big Data in Deep Learning Mode [J]. Journal of Jilin University (Information Science Edition), 2025, 43(2): 231-237.

[10] 王祥,谭国真,彭衍飞,等. 基于语言推理和认知记忆的自动驾驶决策模型 [J]. 吉林大学学报(工学版), 2025, 55(12): 3918-3927.
WANG X, TAN G Z, PENG Y F, et al. Autonomous Driving Decision-Making Model Based on Linguistic Reasoning and Cognitive Memory [J]. Journal of Jilin University (Engineering and Technology Edition), 2025, 55(12): 3918-3927.

[11] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models [J/OL]. 2023 [2026-05-04]. <https://arxiv.org/abs/2307.09288>.

[12] 袁满,张维昱,李明轩. 基于认知图谱的智能问答系统推理模型研究 [J]. 吉林大学学报(信息科学版), 2021, 39(5): 589-595.

YUAN M, ZHANG W G, LI M X. Research on Reasoning Model of Intelligent Q&A System Based on Cognitive Mapping [J]. Journal of Jilin University (Information Science Edition), 2021, 39(5): 589-595.

[13]HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-Rank Adaptation of Large Language Models [J]. International Conference on Learning Representations, 2022, 1(2): 3-10.

[14]PUERTO H, CHUBAKOV T, ZHU X, et al. Fine-Tuning with Divergent Chains of Thought Boosts Reasoning through Self-Correction in Language Models [J/OL]. 2024 [2026-05-04]. <https://arxiv.org/abs/2407.03181>.

[15]赵东, 肖承文, 郭玉庆, 等. 测井软件程序库资源化与运行时部署集成框架 [J]. 吉林大学学报(信息科学版), 2024, 42(5): 972-978.

ZHAO D, XIAO C W, GUO Y Q, et al. An Integrated Framework for Logging Software Library Resourcing and Runtime Deployment [J]. Journal of Jilin University (Information Science Edition), 2024, 42(5): 972-978.

[16]GROENEVELD D, BELTAGY I, WALSH P, et al. OLMo: Accelerating the Science of Language Models [J/OL]. 2024 [2026-05-04]. <https://arxiv.org/abs/2402.00838>.

[17]BIGSCIENCE WORKSHOP, SCAO T L, FAN A, et al. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model [J/OL]. 2022 [2026-05-04]. <https://arxiv.org/abs/2211.05100>.

[18]ZHANG H, ZENG P, GAO L, et al. MPT: Multi-Grained Prompt Tuning for Text-Video Retrieval [C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM, 2024: 1206-1214.

[19]GAO X, XIE W, XIANG Y, et al. Falcon: Faster and Parallel Inference of Large Language Models through Enhanced Semi-Autoregressive Drafting and Custom-Designed Decoding Tree [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Washington DC: AAAI Press, 2025, 39: 23933-23941.

[20]TEAM G, MESNARD T, HARDIN C, et al. Gemma: Open Models Based on Gemini Research and Technology [J/OL]. 2024 [2026-05-04]. <https://arxiv.org/abs/2403.08295>.

[21]JIANG A Q, SABLAYROLLES A, ROUX A, et al. Mixtral of Experts [J/OL]. 2024 [2026-05-04]. <https://arxiv.org/abs/2401.04088>.

(责任编辑: 张洁)