

doi:10.12068/j.issn.1005-3026.2025.20240187

面向多能源发电领域的微调大语言模型 EcoPowerGPT

覃文军¹, 郭彦良^{1,2}, 曲睿婷³, 宋青³

(1. 东北大学 计算机科学与工程学院, 辽宁 沈阳 110169; 2. 应急管理部沈阳消防研究所, 辽宁 沈阳 110034;

3. 国网辽宁省电力有限公司, 辽宁 沈阳 110004)

摘要: 针对多能源发电领域因缺乏高质量数据集导致问答(question answering, QA)效果欠佳的问题, 以及中文回答泛化能力不足的现状, 提出了一种基于Llama架构、面向多能源发电领域的微调大语言模型EcoPowerGPT. 通过整理多能源发电领域的文献与报告, 采用分类过滤和多维度评分方法进行数据处理, 进而构建多能源发电微调数据集, 并基于该数据集对大语言模型进行指令微调. 将EcoPowerGPT在多能源发电QA测试集及单选题测试集上与其他6个对话模型进行对比实验. 结果表明, EcoPowerGPT在回答的准确性与全面性上均优于现有对话模型.

关键词: 生成式大语言模型; 问答; 自然语言处理; 多能源发电; 指令微调

中图分类号: TP 391.1 文献标志码: A 文章编号: 1005-3026(2025)11-0037-11

Fine-Tuned Large Language Model EcoPowerGPT for Multi-energy Power Generation Field

TAN Wen-jun¹, GUO Yan-liang^{1,2}, QU Rui-ting³, SONG Qing³

(1. School of Computer Science & Engineering, Northeast University, Shenyang 110169, China; 2. Shenyang Fire Science and Technology Research Institute of MEM, Shenyang 110034, China; 3. State Grid Liaoning Electric Power Co., Ltd., Shenyang 110004, China. Corresponding author: TAN Wen-jun, E-mail: tanwenjun@cse.neu.edu.cn)

Abstract: To address the issues of poor question answering (QA) performance due to the lack of high-quality datasets in the multi-energy power generation field, as well as the current limitations in the generalization capability of Chinese responses, a fine-tuned large language model called EcoPowerGPT based on the Llama architecture was proposed for the multi-energy power generation field. By organizing literature and reports in the multi-energy power generation field, the model employed classification filtering and multi-dimensional scoring methods for data processing, thereby constructing a fine-tuned dataset for multi-energy power generation. This dataset was then used to fine-tune the large language model. Comparative experiments were conducted between EcoPowerGPT and six other dialogue models on multi-energy power generation QA test sets and test sets of multiple-choice questions with a single correct answer. The results demonstrate that EcoPowerGPT outperforms existing dialogue models in terms of both the accuracy and comprehensiveness of its responses.

Key words: generative large language model; question answering; natural language processing; multi-energy-power generation; instruction fine-tuning

在全球能源需求快速增长和实现可持续发展目标的重要发展方向. 多能源发电技术涵盖火力发电、水力发电、风力发电、核能发电和光伏发电等

收稿日期: 2024-10-21

基金项目: 国家电网有限公司总部科技项目(5108-202218280A-2-404-XG); 国家消防救援局科技计划项目(2025XFZD22).

作者简介: 覃文军(1983-), 男, 山西交城人, 东北大学教授, 博士生导师.

多种能源形式,通过多种能源的互补与协同,构成了现代电力系统的核心.随着全球对环境保护和碳排放减少的需求不断加强,多能源发电不仅能够提升供电系统的稳定性与灵活性,还可以有效减少对单一能源的依赖,降低碳排放强度.特别是清洁能源的引入,通过与传统能源的优化组合,能够进一步推动全球向低碳经济转型,实现更为清洁、绿色的可持续发展.因此,如何提高多能源系统的协同效率、降低整体碳排放,已成为推动未来能源变革的关键.

随着人工智能技术的迅猛发展,自然语言处理(natural language processing, NLP)领域正在经历一场前所未有的变革,而大语言模型的崛起正是这场变革的核心驱动力^[1].从早期的 GPT 系列模型到如今的各种大语言模型,这些参数规模庞大的语言模型凭借深度学习和对海量文本数据的精细训练,具备了令人惊叹的文本生成和理解能力^[2].与传统的语言模型相比,大语言模型不仅能够生成连贯且逻辑严谨的文本,还能在各种复杂的任务中展现出超越以往的智能表现^[3].无论是专业领域的问答、创意写作,还是复杂的语言推理,这些模型都展现出了接近人类甚至超越人类的能力.这些技术的进步彻底改变了我们对机器语言理解能力的传统认知.随着大语言模型的不断优化,它们在多个行业的应用也变得更加广泛,如智能文档处理^[4]、情感分析^[5]、医疗诊断支持等.这不仅提升了工作效率,还推动了企业和社会在数字化转型中的加速发展^[6].

尽管近年来开源的大语言模型和通用模型在自然语言处理方面取得了显著进展,但其在高度专业化领域的表现仍存在较大差距^[7].相较于英语语料库的丰富性和可访问性,中文语料库常常因为数据封闭、质量不均和缺乏公开评审而受限.这种现象使构建高质量的中文预训练模型变得更加困难,同时也限制了中文大语言模型在实际应用中的广泛性和精度.尤其是在一些特定领域,例如多能源发电领域,通用模型由于缺乏领域相关的高质量数据,无法提供准确的技术分析或深入的决策支持.

在多能源发电这类复杂且专业性强的领域,即便是 ChatGPT 这样的先进模型也常面临挑战.该领域技术体系与术语相对复杂,模型需具备精准背景知识,但现有模型往往因领域数据不足难以输出高水平回答.高质量中文数据的稀缺使其在处理相关问题时泛化能力不足,对特定术语的

理解及上下文逻辑处理能力欠缺.这些问题制约了大语言模型在多能源发电领域的实际应用,无法满足行业专业需求.

在多能源发电这一复杂专业领域,现有模型如 ChatGPT 面临领域数据不足、高质量中文数据稀缺等问题,导致回答水平有限、泛化能力弱,难以满足专业需求.本文提出 EcoPowerGPT,通过系统化数据处理获取高质量数据集,结合低秩适应(low-rank adaptation, LoRA)参数高效微调技术进行模型构建.实验表明, EcoPowerGPT 在多能源领域问答效果优于 ChatGPT-3.5 等模型.

为解决多能源发电领域缺乏专用问答数据与适配大模型的问题,本文开展如下研究工作:

1) 构建领域 QA 数据集.为后续模型训练提供优质数据支撑,本文系统收集多能源发电领域论文、指南、报告等资料,通过“能源相关性判别+子领域多分类”的两阶段层次化分类法,结合“领域专业性、问答准确性、可解释性”的多维度评分标准处理数据,构建适配该领域的中文 QA 数据集.

2) 开发领域专用大语言模型.为让模型充分适配多能源发电领域的知识与问答逻辑,以 Llama 3.1-8B 为基座模型,采用 LoRA 微调技术,开发针对该领域的专用大语言模型 EcoPowerGPT.

3) 开展模型性能分析.为验证所开发模型的领域适配效果,计划在多能源发电 QA 数据集(采用 BLEU-4, ROUGE 等指标)与单选题数据集(采用 Accuracy 指标)中,将其与基线模型进行对比分析,以明确其在领域问答任务中的表现.

1 多能源发电大语言模型构建

本文通过整理多能源发电领域的文献与报告,提出系统化的数据处理方法,从大规模多源数据中筛选高质量的能源碳排放相关问答对,并结合参数高效微调技术实现大语言模型在多能源发电领域的高性能表现.多能源发电大语言模型的整体构建思路如图 1 所示.

首先,进行多源数据采集与初步处理,广泛收集多能源发电领域的文献、报告等数据,通过正则匹配、立场识别及数据标准化等操作完成初步处理,随后借助 ChatGPT-3.5 生成领域问答对,为后续流程提供基础数据支撑.

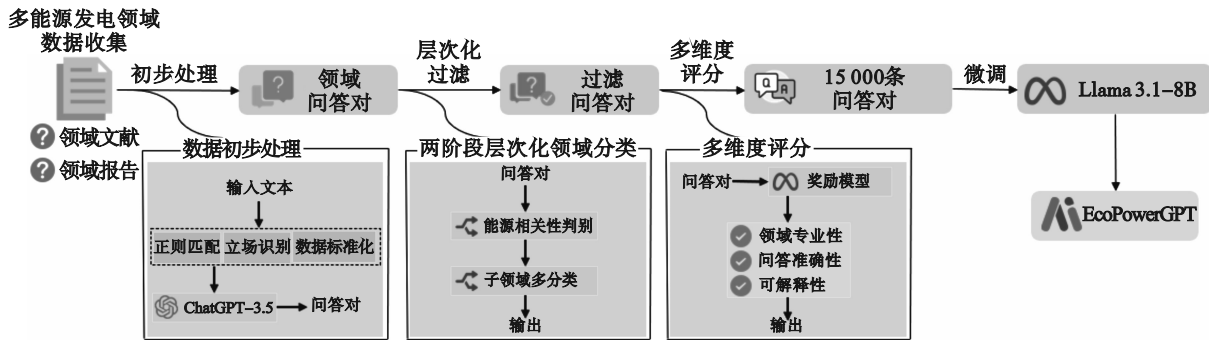


图 1 EcoPowerGPT 构建流程

Fig. 1 EcoPowerGPT construction process

其次,采用两阶段层次化领域分类过滤:第 1 阶段进行能源相关性判别,筛选出与能源领域强相关的问答对;第 2 阶段开展子领域多分类,进一步将问答对细分至具体能源子领域,确保数据的领域聚焦性与专业性.经过层次化过滤后,保留符合要求的问答对进入下一环节.

然后,实施多维度评分筛选,基于领域专业性、问答准确性、可解释性 3 个核心维度构建奖励模型,对过滤后的问答对进行量化评分,最终精选出 15 000 条高质量问答对.该过程通过严格的多维评估,保障数据在内容深度与逻辑严谨性上满足模型训练需求.

最后,进入模型微调阶段,以 Llama 3.1-8B 为基座模型,利用筛选出的高质量问答对进行参数高效微调,使模型充分学习能源领域知识与问答逻辑,最终构建出多能源发电大模型,该方法通过多层次数据采集、系统化过滤机制及精准微调策略,有效提升数据质量与模型性能,确保模型在多能源发电领域实际应用中实现高效、准确的问答交互,为专业场景提供坚实的技术支撑.

1.1 多能源发电领域数据获取与处理

本文对多能源发电数据进行收集与整理,主要包括风能、太阳能、核能等不同发电技术的研究文献和行业报告.这些文献来源于最新的报告、学术论文及政策文件,提供了丰富的领域知识基础,涵盖了设备规格、技术应用、环境影响等方面.对获取到的数据首先通过正则表达式匹配与数据标准化统一单位进行初步处理,并基于大模型构建问答对数据.

由于多能源发电领域涉及风能、光伏、水力、核能及火力等多种子领域,文本数据来源广泛且噪声较大.如何在微调大语言模型时有效过滤与评估数据质量,是提升模型专业性和鲁棒性的关键.本文围绕“领域分类过滤-多维度评分”构建

数据处理流程,提出一套适配该领域的具体数据处理方法,为该领域大语言模型微调提供技术参考.

通过数据处理,本文形成了 1 个包含 15 000 条高质量问答对的多能源发电中文 QA 数据集.

1.1.1 两阶段层次化领域分类

在多能源发电领域,大语言模型的数据处理需兼顾数据质量与领域专业性.本方法通过两阶段层次化分类为大语言模型提供高质量、细粒度的预处理数据,使其在微调后能更好地理解能源领域语义与子领域特征.

第一阶段进行能源相关性判别.由于 BERT 模型在自然语言处理任务中表现出强大的语义表征能力,能够有效捕捉文本的深层语义信息^[8],本文采用开源 bert-base-uncased 编码器,对经清洗与 WordPiece 分词^[9]后的文本 x 输出 [CLS] 向量,如下式(1)所示:

$$h_{\text{CLS}} = \text{BERT}([\text{CLS}], x, [\text{SEP}]) \in \mathbb{R}^d. \quad (1)$$

式中: h_{CLS} 是 BERT 模型处理输入序列后,输出的 [CLS] 标记对应的向量; $\text{BERT}(\cdot)$ 是 BERT 模型的编码函数; [CLS] 是 BERT 模型中的特殊起始标记; [SEP] 是 BERT 模型中的特殊结束标记; d 为 BERT 模型输出的 [CLS] 向量维度.

这一向量富含句子级别的语义信息,适合下游判别任务.由于预训练的 BERT 模型是基于通用语料库进行训练的,虽然具备强大的语义理解能力,但对于能源领域这种专业性较强的领域,其输出的特征可能无法完全适配该领域的特定需求.因此,需要结合后续的分类任务对 BERT 进行微调,使其能够更好地捕捉能源领域文本的特征.

第二阶段进行相关性判断,相关性判别公式如式(2)所示:

$$p_{\text{rel}}(x) = \sigma(\mathbf{w}_{\text{rel}}^T \mathbf{h}_{\text{CLS}} + b_{\text{rel}}). \quad (2)$$

式中: σ 为 Sigmoid 函数,它能够将输入映射到

(0,1)区间,输出一个概率值,便于判断文本与能源领域相关的可能性;在大语言模型数据处理中,剔除数据中混入的、不属于能源领域的文本内容可避免模型学习到无效信息,提升后续微调效率与效果。 $w_{\text{rel}} \in \mathbf{R}^d, b_{\text{rel}} \in \mathbf{R}$ 为可学习参数,通过训练可以优化这些参数,使模型更准确地判断相关性。

本文设置 $\tau_{\text{rel}}=0.5$ 作为默认判定阈值.该值来源于二分类模型常用的“概率中点”策略,能在多数场景下平衡精度与召回率.若 $p_{\text{rel}}(x) < \tau_{\text{rel}}$,则将 x 视为与能源主题无关的样本并直接丢弃;这样可以有效剔除噪声数据,减少无效数据对后续子领域分类的干扰.否则进入子领域多分类环节,确保进入下一阶段的文本都与能源领域相关,从而提高子领域分类的针对性。

在子领域多分类中,本文定义5个子领域标签集 $\{C_1, C_2, \dots, C_5\} = \{\text{风能, 光伏, 水力, 核能, 火力}\}$,覆盖多能源发电核心子领域.对每个“已判定为相关”的文本 x ,使用分类头参数 $\{w_i, b_i\}_{i=1}^5$,通过Softmax函数计算其属于第 i 类的概率,如下式(3)所示:

$$p_i(x) = \frac{\exp(w_i^T h_{\text{CLS}} + b_i)}{\sum_{j=1}^5 \exp(w_j^T h_{\text{CLS}} + b_j)}, i=1, 2, \dots, 5. \quad (3)$$

式中: w_i 为第 i 个子领域分类对应的权重向量; w_j 为对所有子领域的权重向量的遍历表示; b_i 为第 i 个子领域分类对应的偏置项; b_j 为对所有子领域的偏置项的遍历表示。

Softmax函数能够将模型输出转化为概率分布,方便确定文本属于各子领域的概率,从而实现多分类任务.本文损失函数 $L_{\text{dom}}(x)$ 采用加权交叉熵,如下式(4)所示:

$$L_{\text{dom}}(x) = -w(x) \sum_{i=1}^5 y_i(x) \ln p_i(x). \quad (4)$$

其中 $y_i(x)$ 为文本的One-hot标签, $w(x)$ 为术语加权函数.该损失在参数学习中,与BERT模型主干一同反向更新,对其参数进行微调,可以使BERT模型更好地适应能源领域的文本特征,从而提高整个分类系统的性能.引入加权交叉熵的目的是为了突出含有专业术语的文本的重要性,因为在能源领域中,包含专业术语的文本往往更能体现其所属子领域的特征。

本文预先构建能源领域专业术语集 \mathcal{D} ,包含“涡轮机”“光伏组件效率”等关键词,并通过式(5)计算文本中专业术语的密集度。

$$\rho(x) = \frac{|\mathcal{D} \cap \text{Tokens}(x)|}{|\text{Tokens}(x)|}. \quad (5)$$

式中: $\text{Tokens}(x)$ 为文本 x 分词后的所有token构成的集合。

密集度反映了文本中专业术语的占比情况,占比越高,说明文本与能源领域的专业性越强.本文设计的术语加权函数如式(6)所示,主要基于监督式术语加权方法,可以有效放大专业度高的文本的梯度贡献。

$$w(x) = 1 + \eta \rho(x), \eta > 0. \quad (6)$$

式中, η 为调节术语加权强度的超参数。

通过这种方式,当文本中包含较多专业术语(即 $\rho(x)$ 较大)时, $w(x)$ 会相应增大,使得模型在训练过程中更加关注这些专业度高的文本,从而提高模型对能源领域各子领域特征的学习效果,提升分类的准确性。

1.1.2 多维度评分

本文构建了包含5000条人工标注样本的多能源发电数据集,并基于该数据微调Llama 2-7B模型得到奖励模型,用于对多能源发电领域问答数据进行高质量筛选.模型从3个核心维度对每条样本进行评分,分别是领域专业性、问答准确性、可解释性,并通过线性加权筛选出总分大于等于0.6的高质量样本,以确保最终训练集的精炼与可靠。

首先,领域专业性评分 S_{domain} 主要评估回答中行业术语与概念的使用是否恰当、专业.使用微调后的模型对输入回答计算自回归(next-token)、负对数似然(negative log-likelihood, NLL),并将其线性映射到 $[0, 1]$ 区间,计算公式如下式(7)所示:

$$S_{\text{domain}} = 1 - \frac{\text{NLL}(x)}{C}. \quad (7)$$

式中: $\text{NLL}(x)$ 为输入文本 x 对应的负对数似然值; C 为训练集中所有样本NLL的最大值,以保证 $S_{\text{domain}} \in [0, 1]$ 。

然后,问答准确性评分主要用于衡量模型回答与专家参考答案的语义贴合度,将生成回答与参考答案分别编码为向量,并计算它们的余弦相似度 S_{acc} ,如式(8)所示。

$$S_{\text{acc}} = \cos(\mathbf{E}_{\text{pred}}, \mathbf{E}_{\text{ref}}) = \frac{\mathbf{E}_{\text{pred}} \cdot \mathbf{E}_{\text{ref}}}{\|\mathbf{E}_{\text{pred}}\| \|\mathbf{E}_{\text{ref}}\|}. \quad (8)$$

式中: \mathbf{E}_{pred} 为模型生成回答经过编码后得到的向量; \mathbf{E}_{ref} 为专家参考答案经过编码后得到的向量。

可解释性评分 S_{exp} 用于判断模型评分决策是否具备一定透明度.选取回答中若干关键词或句法结构(如因果句),对其进行简单扰动(如增删

关键词),再比较扰动前后模型输出的评分差异,如式(9)所示:

$$S_{\text{exp}} = 1 - \frac{1}{M} \sum_{p=1}^M |r(x) - r(x'_p)|. \quad (9)$$

式中: $r(\cdot)$ 表示单一评分头输出(取值于 $[0, 1]$); x'_p 为对第 p 个位置做微小修改后的输入; M 表示位置个数,为5;此时 $S_{\text{exp}} \in [0, 1]$,数值越大表明评分对小改动越不敏感,可解释性越好.

最后,进行综合加权与数据筛选,总体评分采用线性加权,如下式(10)所示:

$$S_{\text{total}} = \alpha S_{\text{domain}} + \beta S_{\text{acc}} + \gamma S_{\text{exp}}, \alpha + \beta + \gamma = 1. \quad (10)$$

式中:本文权重参数设置为 $\alpha=0.4, \beta=0.4, \gamma=0.2$.针对权重设置,由于多能源发电领域高度依赖专业术语和技术细节,且模型答案必须与专家参考高度一致,故将领域专业性与问答准确性置于同等且较高权重.虽然可解释性对数据筛选的透明度与可追溯性非常重要,但在首轮筛选中,其作用略低于前两者,因此赋予略低权重.

本文将样本总分阈值定为 $\tau=0.6$.原因是三维评分均匀分布在 $0 \sim 1$ 的情形下, $\tau=0.6$ 相当于各维度平均分达到0.6,确保样本在专业性与准确性上均不低于中等水平.同时,本文通过对5 000条样本进行初步打分及人工抽检, $\tau=0.6$ 可筛出约60%的高质量样本,既保证数据量,又良好提升了平均分布和下游训练效果.最终,保留集合 $\mathcal{D}_{\text{good}}$ 定义为:

$$\mathcal{D}_{\text{good}} = \{x \mid S_{\text{total}}(x) \geq 0.6\}. \quad (11)$$

1.2 基座模型

本文构建多能源领域大模型所使用的基础模型是Llama 3.1-8B.Llama 3.1是MetaAI在其Llama系列中的重要更新,基于Llama 3的设计,通过更细致的训练策略进一步提升了模型性能和效率.该模型延续了仅解码器(decoder-only)的Transformer架构,增强了词汇表的容量,从而从原来的32 K提升至128 K,极大地提升了编码效率,使其在处理复杂语言任务时表现得更加游刃有余.此外,Llama 3.1继续使用分组查询注意力机制,有效地优化了推理过程,降低了计算成本,并确保在处理长文本时保持高效与准确性^[10].

本文多维度评分构建的奖励模型,所用基础模型是Llama 2-7B.该模型采用了Transformer架构,并针对多样化的语言任务进行了特别优化.其核心特点之一是相对较小的词汇表(32 K token vocabulary),尽管如此,Llama 2依然在多种

基准测试中展现了卓越的性能^[11].

1.3 大语言模型微调

微调是大语言模型训练的常用技术,在预训练模型基础上调整参数,使其适配特定任务.指令微调是常见策略,通过提供任务指令示例,助力模型学习任务执行方法.此外,我们还可以使用全微调(full fine-tuning, FFT)更新模型全部权重,但是资源消耗较大;使用参数高效微调(parameter-efficient fine-tuning, PEFT)方法只更新模型的一小部分参数^[12],避免了全量微调时的大规模资源消耗,同时保留了模型原有的知识和能力.这种方法提高了内存利用效率并减少了过拟合和灾难性遗忘的风险.其中最具代表性的是2021年微软提出的LoRA方法^[13].之后陆续提出了很多参数高效微调方法,如斯坦福大学提出的Prefix-Tuning^[14]、谷歌提出的Prompt Tuning^[15]、2023年华盛顿大学提出的QLoRA^[16]以及2024年英伟达提出的DLoRA^[17].

本文采用LoRA方法,通过减少可训练参数数量实现低成本高效的模型适配.该方法在预训练模型基础上添加新路径,引入降维矩阵 A 和升维矩阵 B 相乘模拟权重更新,同时“冻结”预训练模型权重,仅训练 A 和 B .

具体而言,设预训练模型某层原始权重为 W_0 ,LoRA通过增加低秩增量 ΔW 调整原有权重,构造方式为引入两个低秩矩阵 A 与 B ,公式如下:

$$\Delta W = BA. \quad (12)$$

式中: $A \in \mathbf{R}^{d \times r}$ 为降维矩阵; $B \in \mathbf{R}^{r \times d}$ 为升维矩阵; r 为低秩维度(通常远小于 d).

因此,经过LoRA微调后的权重可以表示为

$$W = W_0 + BA. \quad (13)$$

对应地,对于输入向量 x ,该层的输出为

$$y = W_0 x + BAx. \quad (14)$$

这两个矩阵的乘积 BA 可直接与预训练模型原始权重叠加,在推理时无需额外计算开销,即可实现与全量微调相当的效果.

2 实验与分析

2.1 实验环境与实验设置

本实验环境配置如下:GPU为NVIDIA RTX 4090;显存为24 GB;Python版本为3.10.8;CUDA版本为12.1;PyTorch版本为2.0.0.本实验大模型LoRA微调参数设置如表1所示.

表 1 大语言模型微调参数设置
Table 1 Fine-tuning parameter settings for large language models

参数	中文含义	参数值
per_device_train_batch_size	每个训练设备上训练的批次大小	2
gradient_accumulation_steps	梯度积累步数	8
max_steps	指定训练的最大步数	5 000
save_steps	模型保存间隔	1 000
learning_rate	学习率	0.000 1
lr_scheduler_type	学习率曲线	cosine
num_train_epochs	训练周期数	4

2.2 基线模型

为了全面评估 EcoPowerGPT 在多能源发电领域问答的有效性,本文选取了 6 款在问答任务中具有代表性且覆盖不同架构特点的主流模型进行对比.这些基线模型既包括完全开源、可自由微调的研究型模型,也涵盖了拥有成熟商业化服务的闭源平台;既有侧重中文处理优化的本土化方案,也有在推理能力上表现突出的轻量化模型.通过在同一评测框架中对这 6 款模型展开横向对比,可以分析 EcoPowerGPT 在多能源发电领域问答能力的优势.所选择的基线模型具体如下:

1) Llama 2 由 Meta 于 2023 年发布^[18],相比第 1 代 Llama,其在推理、文本生成和对话系统等任务上均有显著提升,且支持自定义微调和扩展,已在多个研究和工业场景中展现出了出色的可扩展性^[19].

2) Qwen 2 是阿里巴巴最新推出的开源大模型系列^[20],采用混合专家架构,结合大规模预训练与针对性微调,并行优化模型能力.该系列在多语言理解、文本生成、数学运算与复杂推理等方面表现突出^[21],社区活跃且易于二次开发和应用创新.

3) ERNIE-3.5 是百度推出的商用级大语言模型,综合能力超越 ChatGPT-3.5,并引入插件系统以便接入实时信息和处理长文本;同时结合知识图谱与自适应混合并行训练技术,在逻辑推理和数学计算等复杂任务上具备明显优势.

4) Baichuan 2 是百川智能推出的新一代开源大模型,该模型专注于中文处理和多维度任务表现^[22],在中文处理上展现出显著的优势,其开源特性尤其适合中文开发者^[23].

5) Mistral-7B 模型参数量为 70 亿,通过创新架构在小参数规模下实现了对推理、数学运算和代码生成的卓越性能,同时在推理速度和长序列处理能力上也颇具竞争力^[24].

6) ChatGPT-3.5 作为 OpenAI InstructGPT 的改进版本,专注于对话生成与上下文理解,具备成熟的应用程序编程接口(application programming interface, API)及“安全提示”策略,能生成流畅连贯且逻辑清晰的文本,在智能助手、客服机器人等应用中被广泛采用^[25].

2.3 实验数据

本文所用微调数据集为自主收集并处理的多能源发电领域 QA 数据集,总量 15 000 条,其中水力发电 4 998 条、火力发电 5 007 条、光伏发电 4 994 条、核能发电 4 995 条、风力发电 5 006 条.测试基准包含两个数据集:其一是基于多能源发电报告及学术论文构建的 QA 测试集(含 1 000 个问答对),用于比对输出结果与真实答案;其二是从多能源题库整理的单选题集(含 1 000 题),每题设多个候选答案且仅有一个正确选项,用于评估选择准确性.

数据组织方面,QA 类数据均按子领域分类存储,每条记录包含子领域缩写、序号、问题、答案、来源及子领域标签;训练集额外标注质量评分,测试 QA 集增设真实答案字段,单选题集则包含题干、候选答案与正确选项.采用 JSON 格式存储方案,并对高频访问的测试数据实施内存缓存策略以保障实验连贯性与硬件利用率.

训练集与测试集均源自权威渠道,各子领域数据占比严格控制在 18%~22%,并通过数据溯源确保样本唯一性.训练集侧重 QA 知识学习,测试集通过“答案匹配与单选题选择”多任务模式评估模型泛化能力,有效规避过拟合风险.

2.4 评价指标

针对问答数据集,本文采用 BLEU-4, ROUGE-1, ROUGE-2 和 ROUGE-L 共 4 种评价指标来评估模型性能;针对单选题测试数据集,本文采用 Accuracy 评价指标进行评估,评价指标如下所示:

1) ROUGE是由Lin等^[26]提出的自动化文本摘要评估方法,是基于 N -gram召回率的自动化评估指标,用于度量生成文本与参考文本之间的重叠程度,得分范围为0至1.

ROUGE- N 通过计算生成文本与参考文本在 N -gram级别上的重叠召回率,衡量摘要或答案生成质量.本文选用ROUGE-1和ROUGE-2,分别评估模型在词(1-gram)和短语(2-gram)层面的覆盖情况.

$$\text{ROUGE} - N = \frac{\sum_{t \in R} \sum_{\text{gram}_N \in \text{gram}_N(t)} \text{Count}_{\text{match}}(\text{gram}_N, s)}{\sum_{t \in R} \sum_{\text{gram}_N \in \text{gram}_N(t)} \text{Count}(\text{gram}_N, t)}. \quad (15)$$

式中: R 为参考文本集合; t 为单个参考文本; s 为生成文本; N 指代 N -gram的长度参数; $\text{gram}_N(t)$ 为参考文本 t 中所有 N -gram的集合; $\text{Count}_{\text{match}}(\text{gram}_N, s)$ 为生成文本 s 与参考文本 t 中匹配的 N -gram数量; $\text{Count}(\text{gram}_N, t)$ 为参考文本 t 中所有 N -gram的总数量.当 $N=1$ 时,即ROUGE-1,评估单词层面的重叠率;当 $N=2$ 时,即ROUGE-2,评估连续双词短语的覆盖度.

ROUGE-L通过最长公共子序列(LCS)计算召回率与精确率,既反映了词汇重叠,又兼顾了文本结构,公式如式(16)所示:

$$\text{ROUGE} - L = \frac{(1 + \lambda^2) R_{\text{lcs}} P_{\text{lcs}}}{R_{\text{lcs}} + \lambda^2 P_{\text{lcs}}}. \quad (16)$$

式中: R_{lcs} 代表最长公共子序列的召回率; P_{lcs} 代表最长公共子序列的精确率; λ 是召回率的权重参数.

2) BLEU-4用于衡量模型生成文本与参考答案在词汇和结构上的相似度,通过统计1-gram至4-gram的精确率并结合长度惩罚因子,防止过短输出获得高分^[27],如下式(17)所示:

$$\text{BLEU} - 4 = \text{BP} \times \exp\left(\frac{\ln(P_1) + \ln(P_2) + \ln(P_3) + \ln(P_4)}{4}\right). \quad (17)$$

式中, P_1, P_2, P_3, P_4 分别代表1-gram到4-gram的精确率.长度惩罚因子BP定义为

$$\text{BP} = \begin{cases} 1, & l_c > l_s; \\ e^{1 - \frac{l_c}{l_s}}, & l_c \leq l_s. \end{cases} \quad (18)$$

式中: l_c 为生成文本长度; l_s 为参考文本长度.

3) Accuracy用于评估构建的多能源发电领域单选题数据集,公式如(19)所示.

$$\text{Accuracy} = \frac{\text{CorrectAnswers}}{1\ 000} \times 100\%. \quad (19)$$

式中,CorrectAnswers代表正确的题目数.

2.5 实验结果与分析

根据构建的多能源发电微调数据集,对EcoPowerGPT模型进行了LoRA微调训练,并在训练过程中记录了每轮的损失值.训练损失曲线如图2所示,该曲线展示了模型在训练过程中的表现.损失值是衡量模型输出与实际结果之间差异的指标,数值越小表明模型输出与实际结果越接近,模型性能越好.图中原始曲线代表原始损失值,其波动较为明显,反映了每一轮训练中损失值的具体变化情况;而平滑曲线是对原始损失值进行平滑处理后得到的,它过滤了部分随机波动,更清晰地展现了损失值随训练步骤的总体变化趋势.

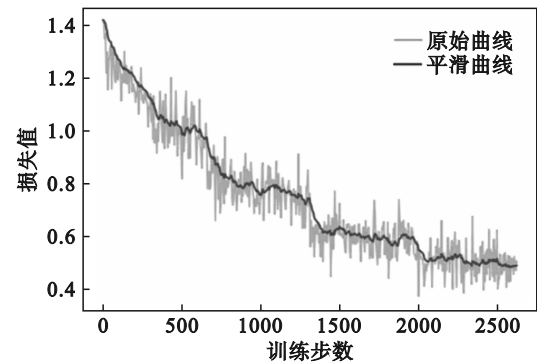


图2 模型训练损失曲线

Fig. 2 Model training loss curves

图2中的曲线清楚地显示,随着训练轮数的不断增加,损失值逐渐降低.这一变化表明,模型逐步学习到了训练集中的特征,并不断改进其预测能力.最终,损失值趋于稳定,表明模型已经达到基本收敛状态,在训练集上的性能也趋于稳定.这种稳定的损失值表明模型在经过充分训练后,其对训练数据的理解和预测能力已趋于成熟.

本文对原始大语言模型与微调后的EcoPowerGPT模型的效果进行了对比评估,结果如表2所示.

表2 原始大语言模型与微调大语言模型对比
Table 2 Comparison between original large language model and fine-tuned large language model

评价指标	原始大语言模型	微调大语言模型
BLEU-4	9.24	61.61
ROUGE-1	34.44	75.17
ROUGE-2	12.19	64.00
ROUGE-L	24.24	70.40

从实验结果可以看出,微调后的大语言模型在BLEU-4, ROUGE-1, ROUGE-2和ROUGE-L

这 4 个指标上都显著优于原始模型。BLEU-4 的提升表明模型在生成文本的准确性方面有了显著进步,而 ROUGE 指标的提升则反映出模型在词汇重叠、连续词组匹配以及文本结构上的表现更加出色。这表明微调后的模型不仅能够更好地生成与标准答案相似的文本,同时在语言流畅性

和语义一致性上也有明显提升,充分验证了微调策略的有效性。

本文基于多能源发电 QA 数据集以及多能源发电领域的单选题数据集将 EcoPowerGPT 与其他基线模型进行比较。首先基于多能源发电 QA 数据集进行测试,结果如表 3 所示。

表 3 多能源发电 QA 数据集评价指标对比

Table 3 Comparison of evaluation indexes for QA dataset of multi-energy power generation

模型	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
EcoPowerGPT	61.61	75.17	64.00	70.40
Llama 2-7B	9.41	27.60	8.92	18.41
Qwen 2-7B	8.15	27.13	9.58	15.73
ERNIE-3.5	29.24	26.61	13.25	14.51
Baichuan 2-7B	6.95	21.78	7.65	14.09
Mistral-7B	5.37	25.95	8.36	14.61
ChatGPT-3.5	44.59	27.85	17.19	15.19

实验结果表明,EcoPowerGPT 在所有评价指标上均表现出显著优势。相比其他模型,EcoPowerGPT 的 BLEU-4 得分高达 61.61,高于其他模型,表明其在多能源发电领域生成与参考文本高度一致的句子方面能力突出。同时,EcoPowerGPT 在 ROUGE-1, ROUGE-2 和 ROUGE-L 得分也分别达到 75.17,64.00 和 70.40,表明其在整体语义结构的一致性上也保持了领先优势。这说明 EcoPowerGPT 在多能源发电领域具备较精准的领域问答能力。

本文基于多能源发电单选题测试集对模型进行测试,结果如表 4 所示。

表 4 多能源发电单选题数据集评价指标对比

Table 4 Comparison of evaluation indexes for dataset of multiple-choice questions with a single correct answer of multi-energy power generation %

模型	Accuracy
Mistral-7B	42.4
Baichuan 2-7B	62.6
Llama 2-7B	64.2
Qwen 2-7B	76.8
ChatGPT-3.5	81.2
ERNIE-3.5	84.2
EcoPowerGPT	90.3

在多能源发电单选题数据集评估中,EcoPowerGPT 以 90.3% 的准确率显著领先于其他基线模型,充分体现了其对多能源发电领域专

业知识的深度理解与语义逻辑的精准把握,表明其在多能源发电场景下具有更全面的知识整合能力和更可靠的推理性能。

本文还针对 EcoPowerGPT 与 Qwen 2-7B 模型在多能源场景下的问答进行了典型问答示例测试对比,如表 5 所示。

在典型问答示例测试中,EcoPowerGPT 展现出高效的推理效率与出色的专业问答能力。相较于 Qwen 2-7B,其回答优势更体现在对问题核心的精准聚焦、专业信息的无冗余传递,以及与多能源领域需求的高度适配。模型回答术语使用精准,逻辑清晰,体现了对能源技术体系的深度理解与知识整合能力。实验结果表明,EcoPowerGPT 在多能源领域问答中,既能快速响应不同复杂度的问题,又能确保回答内容准确专业、逻辑严谨,充分展现了其在实际应用场景中为用户提供高效、可靠知识服务的能力,有效验证了模型在多能源发电领域的实用性。

2.6 模型部署

本文将微调后的 EcoPowerGPT 模型基于 Ollama 框架部署为本地服务,以实现多能源发电领域问答的在线对话能力。Ollama 作为一款开源轻量级框架,可在本地硬件上简化大模型的管理与运行,有效减少对外部 API 的依赖并提升隐私安全性。部署完成后,EcoPowerGPT 既可通过命令行接口直接交互,也可结合轻量级 HTTP 服务以 RESTful API 形式对外提供问答接口,满足电力系统场景中对高并发、低延迟与本地化推理的需求。

表 5 多能源领域 QA 测试对比
Table 5 Comparison of QA testing in multi-energy fields

问题	模型	回答	生成时间
太阳能光伏发电技术与传统能源相比,如何实现能源的清洁和环保	EcoPowerGPT	<p>太阳能光伏发电技术与传统能源相比,实现能源的清洁和环保的主要体现在以下几个方面</p> <ol style="list-style-type: none"> 1. 零排放:光伏发电技术利用太阳能将其转化为电能,不产生二氧化碳等温室气体和污染物的排放,对大气环境没有任何污染 2. 降低碳排放:光伏发电技术的使用减少了对传统能源的依赖,降低了碳排放量,有助于减缓全球气候变暖 3. 减少环境污染:光伏发电技术不需要进行燃烧过程,减少了空气、水和土壤的污染,对环境的影响较小 4. 节约能源资源:光伏发电技术利用太阳能作为能源,不需要消耗有限的化石能源资源,有助于节约能源 5. 降低环境风险:光伏发电技术不需要进行燃烧过程,减少了火灾和爆炸等环境风险 <p>综上所述,太阳能光伏发电技术通过零排放、降低碳排放、节约能源资源、减少环境污染和降低环境风险等方式,实现了能源的清洁和环保</p>	5.23 s
太阳能光伏发电技术与传统能源相比,如何实现能源的清洁和环保	Qwen 2-7B	<p>太阳能光伏发电技术是一种清洁、环保、可持续的能源,它与传统能源相比,实现了以下几点能源的清洁和环保</p> <ol style="list-style-type: none"> 1. 无污染排放:太阳能发电过程不产生任何污染物,包括二氧化碳、二氧化硫、氮氧化物等温室气体和有害气体,因此对大气环境没有污染,有助于减缓全球气候变化和空气污染问题 2. 资源的可再生性:太阳能是一种可再生能源,来源于太阳的光能,理论上可以无限供应,不像煤炭、石油和天然气等化石燃料是有限的资源.使用太阳能发电可以减少对这些有限资源的依赖,促进能源的可持续发展 3. 减少对化石燃料的依赖:传统能源的开采和使用导致了大量的环境破坏和生态失衡.太阳能发电可以减少对化石燃料的依赖,降低对环境的破坏,有助于保护自然资源和生态系统 4. 促进能源结构的多样化:随着太阳能发电技术的不断发展和成本的降低,太阳能成为全球能源结构多样化的重要组成部分.这有助于减少对单一能源的依赖,提高能源系统的稳定性和安全性 5. 提高能源利用效率:现代太阳能发电技术,如高效光伏电池和智能电网系统的集成,可以大大提高能源的利用效率,减少能源浪费,实现资源的高效利用 6. 促进技术创新和就业:太阳能光伏发电技术的发展带动了相关产业的创新和就业,如太阳能电池板制造、系统安装和维护等,为经济发展提供了新的动力 <p>综上所述,太阳能光伏发电技术在实现能源清洁和环保方面具有显著优势,是推动全球能源转型和可持续发展的重要途径.随着技术的进步和成本的降低,太阳能的应用将更加广泛,对全球能源结构和环境保护产生深远影响</p>	10.67 s
《中华人民共和国原子能法(草案)》的出台对核能发展有何作用	EcoPowerGPT	<p>《中华人民共和国原子能法(草案)》的出台可以为核能发展提供法律依据和规范,推动核能产业的健康发展.该法律可以明确核能的合法性和安全要求,加强对核能的管理和监督,提高核能的运营安全性和可靠性</p>	3.49 s

续表 5

问题	模型	回答	生成时间
《中华人民共和国原子能法(草案)》的出台对核能发展有何作用	Qwen 2-7B	《中华人民共和国原子能法(草案)》的出台对核能发展具有多重重要影响	9 s
		<ol style="list-style-type: none"> 1. 法律框架的建立:《中华人民共和国原子能法(草案)》为核能活动提供了一个明确的法律框架,规定了核能活动的合法性和规范性,确保了核能产业的健康发展 2. 安全与监管:通过法律手段,加强了对核能设施和活动的安全监管,确保了核能的使用在安全可控的范围内,保护公众健康和环境安全 3. 促进技术研发与创新:《中华人民共和国原子能法(草案)》鼓励和支持核能技术的研发和创新,推动了核能技术的进步,为核能的高效、清洁利用提供了技术基础 4. 国际合作与交流:该法律可能促进国际间在核能领域的合作与交流,包括技术转让、安全标准的制定与执行,以及核能项目与技术的国际共享 5. 经济与社会影响:《中华人民共和国原子能法(草案)》的实施可能带来经济上的增长,通过核能发电提供稳定的能源供应,减少对化石燃料的依赖,同时创造就业机会,促进相关产业链的发展 6. 环境保护与可持续发展:法律中可能包含对环境保护的严格规定,促进核能产业的可持续发展,减少对环境的影响,支持清洁能源的推广 <p>综上所述,《中华人民共和国原子能法(草案)》不仅为核能产业的发展提供了法律保障,还通过促进安全、技术进步、国际合作、经济与社会影响以及环境保护等多方面,对核能发展产生了深远影响</p>	

在电力系统实际应用中,EcoPowerGPT可直接对接水电、风电、光伏及火电等场站的SCADA系统,通过自然语言接口实时检索各类运行参数(如水轮机出力、风速、面板温度等),并在发现异常时自动生成故障报警与排查建议.调度人员和运维工程师只需以对话方式询问,如“当前风机振动水平是否正常”或“光伏阵列逆变器出现过载怎么办”,即可快速获得巡检指南、故障原因分析及针对不同能源类型的应急预案支持,为水电、风电、光伏等分布式能源场站的高可用性与运行安全提供了有力保障.

3 结 语

本研究构建了针对多能源发电领域的生成式大语言模型 EcoPowerGPT,通过系统化数据处理与高效微调技术显著提升模型专业问答能力.本研究提出两阶段层次化领域分类(能源相关性判别与子领域多分类)及多维度评分的数据处理方法,从多源数据中筛选出 15 000 条高质量中文 QA 数据集,结合 Llama 3.1-8B 基座模型与 LoRA 参数高效微调技术,实现对多能源领域复杂术语的深度学习.实验表明,EcoPowerGPT 在多能源发电 QA 数据集以及多能源发电单选题数据集上

的表现显著优于 Qwen 2, ChatGPT-3.5 等基准模型,有效解决了领域数据不足与中文处理能力薄弱的问题.该模型为多能源领域智能问答系统提供了高性能解决方案,所提出的数据处理方法可为垂直领域大模型构建提供参考,推动相关技术在能源行业的实际应用与研究进展.

同时,本研究仍存在可改进之处.其一,数据以静态文本为主,缺乏多模态信息,知识时效性与场景适配性有待提升;其二,模型对复杂推理与数值计算的支持不足,架构未充分融合领域因果逻辑;其三,实验评估场景较单一,未覆盖实际运维、调度等场景.后续可围绕上述方向补充数据、优化模型架构与评估体系,推动模型向实用落地迈进.

参考文献:

- [1] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. *Journal of Machine Learning Research*, 2020, 21(1): 5485-5551.
- [2] 周昆, 朱余韬, 陈志朋, 等. YuLan-Chat: 基于多阶段课程学习的大语言模型[J]. *计算机学报*, 2025, 48(1): 1-18. (Zhou Kun, Zhu Yu-tao, Chen Zhi-peng, et al. YuLan-chat: a large language model based on multi-stage curriculum learning [J]. *Chinese Journal of Computers*, 2025, 48(1): 1-18.)
- [3] OpenAI, Achiam J, Adler S, et al. GPT-4 technical report

- [EB/OL]. (2023-03-15) [2024-10-19]. <https://arxiv.org/abs/2303.08774>.
- [4] 田萱, 吴志超. 基于信息检索的知识库问答综述[J]. 计算机研究与发展, 2025, 62(2): 314-335.
(Tian Xuan, Wu Zhi-chao. Review of knowledge base question answering based on information retrieval [J]. *Journal of Computer Research and Development*, 2025, 62(2): 314-335.)
- [5] 李诗晨, 王中卿, 周国栋. 大语言模型驱动的跨领域属性级情感分析[J]. 软件学报, 2025, 36(2): 644-659.
(Li Shi-chen, Wang Zhong-qing, Zhou Guo-dong. LLM enhanced cross domain aspect-based sentiment analysis [J]. *Journal of Software*, 2025, 36(2): 644-659.)
- [6] 宫丽娜, 周易人, 乔羽, 等. 预训练模型在软件工程领域应用研究进展[J]. 软件学报, 2025, 36(1): 1-26.
(Gong Li-na, Zhou Yi-ren, Qiao Yu, et al. Research progress of pre-trained model in software engineering [J]. *Journal of Software*, 2025, 36(1): 1-26.)
- [7] Vakili T, Lamproudis A, Henriksson A, et al. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data [C]//Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, 2022: 4245-4252.
- [8] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, 2019: 4171-4186.
- [9] Song X Y, Salcianu A, Song Y, et al. Fast WordPiece tokenization [EB/OL]. (2020-12-31) [2024-10-19]. <https://arxiv.org/abs/2012.15524>.
- [10] Grattafiori A, Dubey A, Jauhri A, et al. The Llama 3 herd of models [EB/OL]. (2024-07-31) [2024-10-19]. <https://arxiv.org/abs/2407.21783>.
- [11] Basile P, Musacchio E, Polignano M, et al. LLaMAntino: LLaMA 2 models for effective text generation in Italian language [EB/OL]. (2023-12-15) [2024-10-19]. <https://arxiv.org/abs/2312.09993>.
- [12] Fu Z H, Yang H R, So A M, et al. On the effectiveness of parameter-efficient fine-tuning [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37(11): 12799-12807.
- [13] Hu E J, Shen Y L, Wallis P, et al. LoRA: low-rank adaptation of large language models [EB/OL]. (2021-06-17) [2024-10-19]. <https://arxiv.org/abs/2106.09685>.
- [14] Li X L, Liang P. Prefix-tuning: optimizing continuous prompts for generation [EB/OL]. (2021-01-01) [2024-10-19]. <https://arxiv.org/abs/2101.00190>.
- [15] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning [EB/OL]. (2021-04-17) [2024-10-19]. <https://arxiv.org/abs/2104.08691>.
- [16] Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: efficient finetuning of quantized LLMs [EB/OL]. (2023-05-23) [2024-10-19]. <https://arxiv.org/abs/2305.14314>.
- [17] Gao C, Zhang S Q. DLoRA: distributed parameter-efficient fine-tuning solution for large language model [EB/OL]. (2024-04-08) [2024-10-19]. <https://arxiv.org/abs/2404.05182>.
- [18] Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models [EB/OL]. (2023-07-19) [2024-10-19]. <https://arxiv.org/abs/2307.09288>.
- [19] Nguyen T T, Wilson C, Dalins J. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts [EB/OL]. (2023-08-28) [2024-10-19]. <https://arxiv.org/abs/2308.14683>.
- [20] Yang A, Yang B S, Hui B Y, et al. Qwen2 technical report [EB/OL]. (2024-07-15) [2024-10-19]. <https://arxiv.org/abs/2407.10671>.
- [21] Zhang S, Peng B C, Zhao X P, et al. LLaSA: large language and E-commerce shopping assistant [EB/OL]. (2024-08-02) [2024-10-19]. <https://arxiv.org/abs/2408.02006>.
- [22] Yang A Y, Xiao B, Wang B N, et al. Baichuan 2: open large-scale language models [EB/OL]. (2023-09-19) [2024-10-19]. <https://arxiv.org/abs/2309.10305>.
- [23] Cao Y Q, Yang L, Wei C, et al. Financial text sentiment classification based on Baichuan2 instruction finetuning model [C]//2023 5th International Conference on Frontiers Technology of Information and Computer (ICFTIC). Qiangdao, 2024: 403-406.
- [24] Jiang A Q, Sablayrolles A, Mensch A, et al. Mistral 7B [EB/OL]. (2023-10-10) [2024-10-19]. <https://arxiv.org/abs/2310.06825>.
- [25] B T B, Chen J M. Performance assessment of ChatGPT versus bard in detecting Alzheimer's dementia [J]. *Diagnostics*, 2024, 14(8): 817.
- [26] Lin C Y. ROUGE: a package for automatic evaluation of summaries [C]//Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2004: 74-81.
- [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 on Association for Computational Linguistics-ACL '02. Philadelphia. 2001: 311-318.