DOI: 10.19783/j.cnki.pspc.240485

基于 BERT 模型的主设备缺陷诊断方法研究

杨 虹，孟晓凯，俞 华，白 洋，韩 钰，刘永鑫

（国网山西省电力公司电力科学研究院，山西 太原 030002）

摘要：主设备缺陷诊断旨在及时定位处理电网的异常情况，是电力系统平稳运行的基础。传统方法以人工为主，存在效率低下、诊断成本高、依赖专家经验等问题。为了弥补这些不足，提出了一种基于 BERT 语言模型的主设备缺陷诊断方法。首先，使用 BERT 初步理解输入，获取嵌入表示，结合缺陷等级分类任务判断故障的危急程度。然后，利用大语言模型汇总输入信息和评判结果，并通过大语言模型提示学习提高知识问答过程的准确性与推理可靠性，返回正确有效的回答。最后，探究了大语言模型在电力领域的应用潜力。实验结果表明，所提方法在缺陷等级分类任务和问答任务上都表现良好，可以生成高质量的分类证据和指导信息。

关键词：缺陷诊断；大语言模型；BERT；提示学习；分类模型

Research on primary equipment defect diagnosis method based on the BERT model

YANG Hong, MENG Xiaokai, YU Hua, BAI Yang, HAN Yu, LIU Yongxin

(State Grid Shanxi Electric Power Research Institute, Taiyuan 030002, China)

Abstract: Primary equipment defect diagnosis aims to promptly locate and address abnormal situations in the power grid, serving as a foundation for the stable operation of the power system. Traditional methods rely heavily on manual efforts, leading to low efficiency, high diagnostic costs, and dependence on expert experience. To overcome these limitations, this paper proposes a primary equipment defect diagnosis method based on language models such as BERT. First, the BERT model is employed to preliminarily comprehend the input and obtain embedded representations, which are then used in the defect level classification task to assess the severity of the defect. Subsequently, a large language model is utilized to consolidate the input information and classification results, improving the accuracy and reasoning reliability of the knowledge-based Q&A process through prompt learning, thereby providing correct and effective answers. Finally, the potential applications of large language models in the power industry are explored. Experimental results demonstrate outstanding performance of this method in both defect level classification and question-answering tasks, generating high-quality classification evidence and guidance information.

This work is supported by the General Program of National Natural Science Foundation of China (No. 62176227).

Key words: defect diagnosis; large language model; BERT; prompt learning; classification method

0 引言

电力系统是现代社会不可或缺的基础设施，它支撑着工业生产、商业运营和日常生活，保障了医疗、通信、信息技术等重要行业的正常运转。面对不断增长的能源需求和可持续发展的挑战，电力系统作为推动社会进步和创新的重要支撑，其重要性愈发凸显。然而，电力系统的设备缺陷可能直接影响其稳定性与可靠性，主设备(如变压器、电抗器等)的缺陷可能导致能源传输中断，影响工业、商业和生活用电。因此，对主设备缺陷进行及时定位和处理，对维持电力系统的正常运行至关重要。这一任务被称为主设备缺陷诊断，其主要目的在于对设备缺陷的情况进行评估，确保电力系统运维人员可以根据设备的实际状况和缺陷的严重性来制定维护计划，确保系统高效运行和长期稳定性。

传统的缺陷诊断主要由人工完成，以规程文档或缺陷手册为基础，耗时耗力，对从业人员的技术水平和相关经验都有所要求。随着智能电网概念的

基金项目：国家自然科学基金面上项目资助(62176227)；国网山西省电力公司科技项目资助(52053023000P)

推广和电力大数据时代的来临，传感监控系统、大数据分析、专家系统、机器学习算法等先后进入缺陷诊断体系，有效提高了从业人员的工作效率[1-3]。文献[4]深入分析过去十余年的缺陷记录数据，从设备、部件、缺陷情况等多个角度挖掘重点设备及家族性缺陷设备，并给出了针对性的分析建议。文献[5]利用改进贝叶斯网络建立故障类型与故障位置之间的映射关系，全面捕捉变压器整体运行状态。文献[6]针对缺陷样本不足的问题，提出了一种混合过采样技术，并借助灰狼优化(grey wolf optimizer, GWO)算法优化支持向量机(support vector machine, SVM)模型，提升了模型的整体性能。文献[7]针对历史数据文件进行知识抽取与整合，形成一套基于知识图谱与变电站配置描述(substation configuration description, SCD)文件的安全措施知识库，直接应用于智能运检。文献[8]将近似动态规划思想引入能源系统调度，提高风险检测的效率，以实现实时的风险规避。

这些方法从多个视角提供了自动化缺陷诊断方案，但在实际应用中，仍存在未解决的共性问题。首先，这些方法都具有输入输出局限，虽然可以通过词嵌入方法加以弥补，但仍无法深入理解不同设备名称、缺陷描述之间细微的语义差距。其次，对运维人员而言，推理结果依赖于推理依据，而机器学习方法大都缺少生成推理依据的能力，无法适应人类认知习惯。最后，多数模型依赖数值检测，在状态含糊、缺少确切信息的情况下，无法发挥作用。文献[9]和文献[10]虽然在文档的向量化表示与知识问答方面进行了一些研究，但本质上仍然是对现有知识库的检索利用，并未涉及更深层的推理。

为了应对这些问题，本文提出了一种基于BERT、ChatGLM 等语言模型的主设备缺陷诊断方法，具体来说，这一种由 BERT[11]和 ChatGLM3 组成的两阶段问答模型。首先，利用 BERT 理解输入信息，挖掘蕴含语义，对缺陷情况进行初步判断。然后，引入对话式大语言模型 ChatGLM3，借助该模型的自然语言理解能力，理解故障现状和由 BERT 模型给出的判断结果，针对故障现象进行分析，并给出解决方案。最后，本文探究了将大语言模型应用在电力领域的其他方法。本文选用分类任务和问答任务评价基于 BERT、ChatGLM 等语言模型的主设备缺陷诊断方法的效果，实验结果表明，本文所提方法能够针对用户提出的问题提供高质量的答案。

1 基于 BERT 的文本分类

BERT 是一种基于 Transformer[12]的自然语言表示模型，与前期经典词嵌入方法，如 GloVe[13]、Word2vec[14]等相比，BERT 引入注意力机制来捕捉相关的语义信息，以学习每个词语的上下文表示。

BERT 的基本结构如图 1 所示，针对输入“油浸变压器渗油”，每个字被视作一个“单词”，其初始表示分为 3 部分：与语义相关的的词嵌入、标记句子的段嵌入和位置嵌入，三者经特征求和获得输入表示。之后利用 Transformer 编码器进行 “遮蔽语言模型”和“下一句预测”两步训练。“遮蔽语言模型”指随机遮蔽一部分输入，要求模型预测被遮蔽的内容；“下一句预测”指随机筛选两个句子 A、B，判断句子 B 是否是句子 A 之后的下一个句子。这两步训练内容都未涉及数据标注，是自监督过程。模型通过自监督学习完成预训练后，针对不同的下游任务，只需要使用少量的数据调整模型参数。

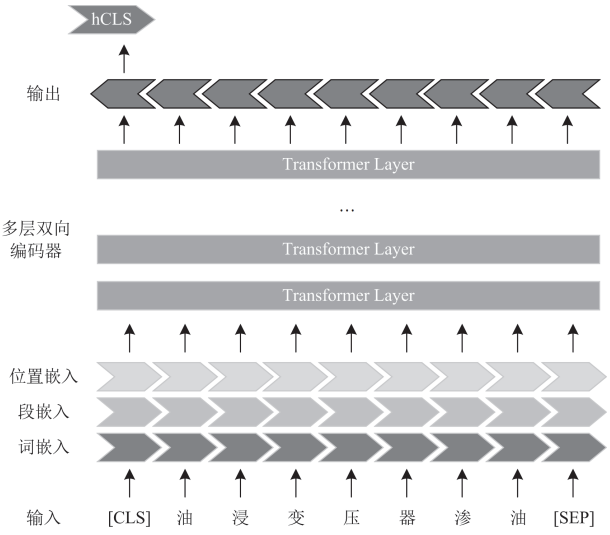

图 1 BERT 模型结构图

Fig. 1 Structure of BERT model

单个 Transformer 层的结构如图 2 所示。原始的 Transformer 模型包括了编码器和解码器，在 BERT 中只使用了编码器部分结构，包括多头注意力、残差连接与归一化、前馈层及第二层残差连接与归一化。

与同量级模型(如 GPT-2)相比，BERT 在中文上具有更佳的表现，本文选择使用由谷歌开源的中文预训练 BERT 模型 BERT-base-chinese，该模型支持简体和繁体中文，可以满足基本的中文理解需求。

文本分类步骤选用 BERT 获取句嵌入，用于下游任务。模型在训练过程中将参数设为可变，即对 BERT 模型进行微调。BERT 模型有两种常用的输出格式，一种被称为“pooler output”，它提取 BERT 中每个句子句首标识符“[CLS]”的嵌入表示；另一种被称为“sequence output”，输出每个字的表示。前者

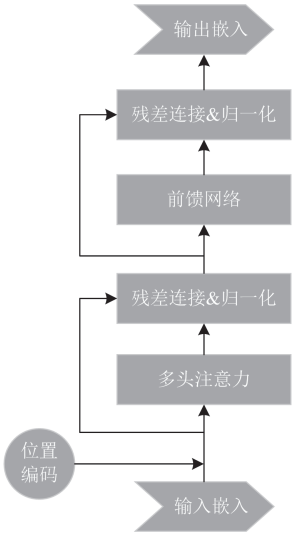

图 2 Transformer 层结构图

Fig. 2 Structure of a Transformer layer

更侧重于句子的表示，适用于文本分类任务，此处采取这种输出，将[CLS]符号的嵌入表示记为 $\boldsymbol{h}_{\text{CLS}}$。

模型在 BERT 基础上增加了一个全连接层，用于承担分类任务。该层的主要功能是将句子的嵌入表示映射到分类空间中，该过程如式(1)所示。

$$\boldsymbol{h}_{\text{output}} = \text{Softmax}(W \cdot \boldsymbol{h}_{\text{CLS}} + b) \tag{1}$$

式中：$\boldsymbol{h}_{\text{output}}$ 为本层的输出；Softmax(·) 表示多分类激活函数；W、b 分别表示该全连接层的权重和偏置。

该过程使用交叉熵损失函数检验模型训练状态，并使用 Adam 优化器进行优化，交叉熵损失函数可表示为

$$L = -\frac{1}{N}\sum_{n=1}^{N}\log(P(n)) \tag{2}$$

式中：L 为损失值；N 为样本数量；$P(n)$ 为序列为 n 的样本分类结果正确的概率，此处为输出向量 $\boldsymbol{h}_{\text{output}}$ 中对应的表示。

根据全连接层的维数设置，可以利用 BERT 完成二分类和多分类任务。本文利用该方法完成两步分类：1) 缺陷等级分类，目的是预估缺陷的严重程度；2) 缺陷依据分类，旨在初步识别可能的分类标准，以帮助大语言模型更精准地进行问答。此外，BERT 模型本身起到了编码器的作用，可以将输入的自然语言转化为字嵌入和词嵌入，允许下游任务将向量化表示引入机器学习方法中，完成更多操作处理。

2　基于大语言模型提示学习的问答增强

大语言模型本身具备了一定的基础知识和领域知识，可以针对不同的问题给出对应的回答[15]。该回答通常具有较好的可读性和一定的逻辑性，但在电力领域，专业术语繁多、知识庞杂，大语言模型可能很难理解领域背景，导致输出有偏差甚至错误的回答，即出现“幻觉”问题。为了提高回答的质量，可以通过变更输出内容的方式引导模型回答问题，即提示工程。本节将从提示的角度入手，采用角色扮演与思维链技术相结合的方法，提高大语言模型在问答任务上的能力。

2.1 提示模板设计

2.1.1 角色扮演方法

角色扮演提示是在与大语言模型正式交互前预先构建会话场景的一系列行为的统称，常包含对场景和身份的预设。大语言模型的训练数据来源于自然语言，一般认为，在自然语言场景中，知识的质量、相关度、可信度往往与其来源密切相关。角色扮演方法是最常见的一类提示，这种提示可能比较简短，只涉及身份，如“你是一名电力工程师”，也可能添加如背景、对话示例等内容。

此处使用角色扮演方法设计提示模板的目的有两点：1) 使回答更贴合电力领域；2) 提高回答的可信度。为此，在角色扮演提示中应重点考虑适应电力领域的要求和回答内容可信。

综合考虑两点需求，角色扮演提示设计如下。

“你是一名电力领域专家。从现在起，你将参加电网专业知识测验，回答各种问题并确保答案的准确性。”

提示包括专业身份信息“电力领域专家”和场景信息“知识测试”，既强调了领域特点，又隐含了知识倾向，即“电网专业知识测试题”，可以覆盖提示需求。该提示将置于提示模板开头。

2.1.2 思维链技术

思维链技术的核心思想是引导大语言模型将一个复杂问题分解为多个子问题，并分步求解，从而提升大语言模型的解题能力[16]。一个完整的思维链提示分指令、逻辑依据、示例三部分。指令用于限定大语言模型的输出格式和输出范围，如“输出推理过程”、“以 json 格式输出回答”。逻辑依据对推理过程进行引导，如可能涉及的推理路径、可能相关的推理步骤。示例是以少样本的方式直接给出一个或多个理想的输入输出，每一个示例都至少包括问题与回答。

思维链提示技术可分为零样本提示和少样本提示，两者的直接区别在于是否给出示例。零样本提示不涉及任何问答示例，也不直接给出专业知识，往往比较简短，如只在原问题之后增加一句“注意一步步思考”。少样本提示则会给出一个或多个示

例，即问题、解答过程、推理路径三者的结合体，引导大语言模型模仿示例问题的解答过程，输出正确答案。思维链提示案例如图 3 所示。

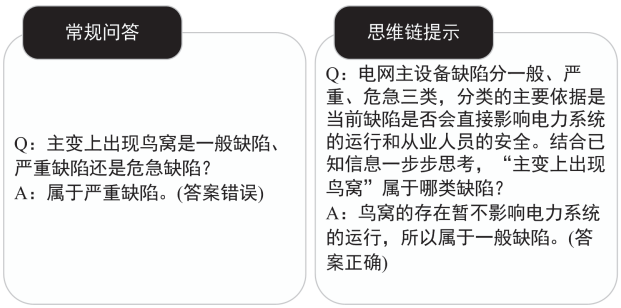

图 3 思维链提示示例图

Fig. 3 An example of chain of thought prompts

主设备缺陷诊断标准多样，如果设计零样本提示，需要提炼一致的业务逻辑，这是难以实现的。因此，本文选择使用少样本提示，通过向大语言模型提供一个推理示例和简单的推理引导的方法，实现思维链提示。

2.2 预处理知识植入

第 1 节已经从缺陷等级、分类依据两个角度对输入问题进行了初步处理，但从提高结果可读性的角度出发，仍需使大语言模型理解整合分类结果。此外，将分类结果融入提示，也有利于大语言模型捕捉问题中的关键点，给出更有针对性的答案。

2.2.1 结果信息直接融合

使大语言模型理解分类结果最直接方法，是将该信息直接插入提示文本中。具体来说，针对缺陷等级分类结果，在填写提示模板时，额外加入“分类结果”一项，并在之后填写分类结果、置信度，如：“分类结果：严重(置信度：0.85)”。置信度数值可由分类器输出直接获得，为增强表现力，此处额外进行了一步换算，若置信度高于 0.7，将标注为“高”，置信度在 0.5~0.7 之间，将标注为“中”，否则标注为“低”。进行换算后，示例将变更为：“分类结果：严重(置信度：0.85，高)”。针对缺陷依据分类结果，同样加入“分类依据”项，直接填写对应分类依据的文本信息。

2.2.2 结果信息辅助提示示例设计

除直接揭示缺陷现象危急程度与缺陷原因外，分类结果还间接起到了案例过滤的作用，这允许模型使用缺陷等级或分类依据作为过滤条件，在训练集中检索与用户输入关联更紧密的示例，用作提示。由于缺陷依据类型繁多，难以准确区分，根据缺陷获取相似案例可能不够可靠，且容易出现过滤结果为空的问题，本文选择缺陷等级进行初步过滤。

考虑到电力设备之间的巨大差距，在进行过滤时，除缺陷等级，额外加入设备类型作为约束，即要求设备类型相同、缺陷等级相同。若过滤结果为空，则变更为只依据缺陷等级进行过滤，优先确保缺陷场景的相似度。为检索更优质的推理路径，除过滤外，此处针对问题描述新增了自监督评估机制。该评估机制基于 K-Means 聚类算法，以句嵌入表示 $\boldsymbol{h}_{\mathrm{CLS}}$ 为基础在每一类训练数据内部构建聚类。K-Means 聚类算法的流程如下。

1) 随机选取 k 个质心；

2) 针对每一个节点，计算其与各个质心的坐标，将其分配到对应的簇；

3) 判断是否有节点的簇发生改变，如果没有，结束迭代，输出聚类结果，否则继续进行步骤 4)；

4) 更新质心坐标，新坐标为该簇内所有节点坐标的平均值；

5) 继续进行步骤 2)。

因缺陷等级主要分为“一般”、“严重”、“危急”三类，此处 k 值初始设定为 3。经过多轮迭代后，模型趋于收敛，可获得聚类质心坐标，可以根据各训练数据与聚类质心之间的距离对其排序，选择 k 个中心节点，其距离 $d_{(w_i,w_t)}$ 计算公式为

$$d_{(w_i,w_t)}=\sqrt{\sum_{u=1}^{D}\left|w_{iu}-w_{tu}\right|^2} \tag{3}$$

式中：D 为聚类空间的维数；w_i 表示当前聚类内第 i 个节点的坐标；w_t 为中心节点的坐标；w_{iu}、w_{tu} 分别为第 i 个节点、中心节点的坐标在第 u 个维度的分量。距离越小，表示该节点越靠近质心，即该案例越典型。

2.3 基于大语言模型提示学习的问答任务

汇总上述内容，大语言模型提示模板共由 5 部分组成：角色扮演提示、提示示例、指令提示、问题、文本分类结果，问题与文本分类结果为嵌套关系。其组成关系如下。

1) 角色扮演提示。预设大语言模型的角色和回答问题的场景。

2) 提示示例。包括一个正确的、由聚类方法筛选获得的问答示例，该事例由问题和答案两部分组成，答案又由缺陷原因和推理路径组成。

3) 指令提示。引出大语言模型需要回答的问题，比较简短，如“需要回答的问题是”。

4) 问题。在用户输入的缺陷现象的基础上，添加引导语句和补充信息获得。

5) 文本分类结果。文本分类结果嵌套在问题中，是补充信息的一部分。

在提示模板的引导下，大语言模型将以半结构化的方式输出答案，允许通过简单的程序实现批量化处理。

3　BERT 模型实验结果与分析

3.1 实验数据

实验原始数据集为电力主设备缺陷故障表。该表来自多地主设备日常运维数据，经人工整理、清洗、去重后获得，表中包括了多种缺陷标准库中涵盖的典型缺陷及不属于标准库的罕见缺陷，其详细信息如表 1 所示。针对缺陷等级分类任务，本文选取该数据中“缺陷现象”与“缺陷等级”两列，构建缺陷等级分类数据集。其中，“缺陷现象”列记录了设备缺陷的表现，是一句自然语言；“缺陷等级”列则记录了缺陷的严重程度。针对缺陷依据分类任务，本文选取该数据中“缺陷现象”与“分类依据”两列，构建缺陷依据分类数据集。其中，“分类依据”是一句在标准库中有记录的自然语言。

表 1 实验数据信息

Table 1 Information of experimental data

缺陷等级	数量	占比/%
一般缺陷	50 831	68.9
严重缺陷	18 414	24.9
危急缺陷	4566	6.2
全部缺陷	73 811	100

在“缺陷等级”分类数据集中，缺陷等级分一般、严重、危急三类。危急缺陷指设备或其他外物出现了直接威胁安全运行、亟需处理的问题，若不及时处理，可能会威胁人身安全，造成设备损坏、大面积停电甚至火灾等事故；严重缺陷指当前缺陷对人身安全或平稳供电有威胁，但尚且不影响设备正常运行，需要尽快处理；一般缺陷影响级别较低，影响范围不大。设备缺陷等级并非一成不变，它可能随着设备运行状态或气候环境的变化而变得更加严重或逐渐消失，把握一般缺陷的演化方向，辨别严重缺陷与危急缺陷的紧迫程度，是缺陷等级分类任务的重点与难点。

在“分类依据”数据集中，统计可得，分类依据共有 549 种，如“缺陷暂不影响运行”、“轻度：无标识或缺少标识”等，这 549 类标签中仅 2 类标签在缺陷依据分类数据集中出现 3500 次以上，即占比超过 5%，最高项占比为 7.1%。多数分类依据出现占比在 0.1%以下，多分类、少样本问题是该分类任务的最主要挑战。

3.2 实验设置

为证明模型的有效性，本文选择了文本分类的常用方法 Transformer[12]、FastText[17]、DPCNN[18]及经典方法 TextCNN[19]、TextRNN[20]进行比较。其中，FastText 使用了经典的连续词袋(continuous bag of words)结构，可应用于多类别、大数据量场景；DPCNN 又称“深层金字塔模型”，通过减采样和残差连接的设计，抽取长文本中词语词之间可能出现的远距离依赖关系；TextRNN 与 TextCNN 则与传统 RNN、CNN 网络相似。TextRNN-Att 表示在原始 TextRNN 的基础上引入了注意力机制，TextRCNN 则表示在 TextRNN 的基础上添加了池化层。

针对 BERT 模型，本文设置随机失活值为 0.5，训练轮数为 100，批大小为 128，短切长度为 32，学习率为 0.000 05，输出维度为 768。本文通过单一线性层实现分类任务，该线性层的输入维度为 768，输出维度受分类数量影响。FastText 模型的隐藏层维度为 256；DPCNN 模型的卷积核数量为 256；TextCNN 模型的卷积核数量为 256；TextRCNN、TextRNN-Att 与 TextRNN 模型的隐藏层维度为 256，隐藏层数量为 1。以上基准模型的学习率均设置为 0.001，其他未提到的参数与 BERT 模型保持一致，或遵循原始论文初始设置。Transformer 模型的隐藏层维度为 1024，最后一个隐藏层的维度为 512，编码器数量为 2，注意力头数为 5，每个注意力头的维度为 300，其余设置与 BERT 模型保持一致。

诊断流程如图 4 所示。整个缺陷诊断流程包括缺陷等级分类、缺陷依据分类、问答 3 个子任务。其中，缺陷等级分类、缺陷依据分类任务由 BERT 模型实现，本文随后将对其实验结果进行重点介绍与分析；问答任务由大语言模型 ChatGLM3[21]实现。

3.3 评价指标

针对缺陷等级分类任务，本文选取准确度 Accuracy，精确率 Precision，召回率 Recall 和 F1 值评价实验结果。这些概念由混淆矩阵获得，分类结果的混淆矩阵如表 2 所示。

设 N_{TP}、N_{FP}、N_{TN}、N_{FN} 分别表示真正例、假正例、真反例、假反例，则准确率 S_{Accuracy}、精确率 $S_{\mathrm{Precision}}$、召回率 S_{Recall}、F1 值 S_{F1core} 表达式分别为

$$S_{\mathrm{Accuracy}}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}} \tag{4}$$

$$S_{\mathrm{Precision}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}} \tag{5}$$

$$S_{\mathrm{Recall}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{6}$$

$$S_{\mathrm{F1core}}=\frac{2\cdot S_{\mathrm{Precision}}\cdot S_{\mathrm{Recall}}}{S_{\mathrm{Precision}}+S_{\mathrm{Recall}}} \tag{7}$$

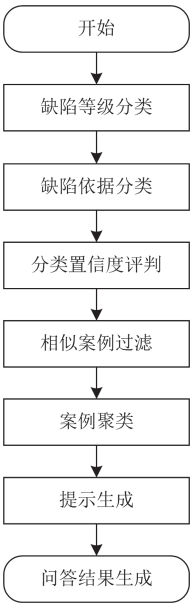

图 4 缺陷诊断流程图

Fig. 4 Defect diagnosis flowchart

表 2 分类结果混淆矩阵

Table 2 Confusion matrix for classification results

真实情况	预测情况	
	正例	反例
正例	真正例(TP)	假反例(FN)
反例	假正例(FP)	真反例(TN)

3.4 实验结果分析

模型在全部数据上的分类结果如表 3 所示。从实验结果可以看出，本文使用的 BERT 模型在多个评价指标都远远超过了基准方法，即能在大部分测试数据上取得正确的结果，具有一定的实用性。

表 3 全部设备实验结果

Table 3 Experimental results of all equipment

方法	准确率	精确率	召回率	F1 值
TextCNN	0.8021	0.7913	0.8021	0.7932
TextRNN	0.7895	0.7785	0.7895	0.7812
TextRNN-Att	0.7928	0.7877	0.7928	0.7888
TextRCNN	0.8084	0.8003	0.8084	0.8019
FastText	0.8055	0.7989	0.8055	0.7998
DPCNN	0.7956	0.7856	0.7956	0.7881
Transformer	0.7754	0.7661	0.7754	0.7612
BERT(Ours)	**0.8730**	**0.8575**	**0.8730**	**0.8627**

模型仅针对主变压器缺陷进行分类的结果如表 4 所示。可见，模型虽仍能取得超过其他基准方法的实验结果，但优势较全部数据集明显偏低。考虑到主变压器是训练数据集中出现频率最高的设备，这可能是因为主变压器这一类设备出现缺陷的次数更多，训练数据集更充分，模型在该类设备上可以得到充分的训练。由该现象可以推知，BERT 模型可以在更少的训练数据上取得更好的训练结果，且不容易出现过拟合问题，这可能与 BERT 模型卓越的语义理解能力相关。与经典文本分类方法相比，本文所提方法可以深入理解文本语义信息，有效建模句与句、词与词之间的关系，在更小的数据集上更有效地捕捉每一类样本的特点。

表 4 主变压器实验结果

Table 4 Experimental results of main transformers

方法	准确率	精确率	召回率	F1 值
TextCNN	0.8640	0.8548	0.8640	0.8528
TextRNN	0.8492	0.8306	0.8492	0.8353
TextRNN-Att	0.8503	0.8378	0.8503	0.8359
TextRCNN	0.8570	0.8469	0.8570	0.8492
FastText	0.8585	0.8525	0.8585	0.8532
DPCNN	0.8492	0.8311	0.8492	0.8300
Transformer	0.8420	0.8339	0.8420	0.8379
BERT(Ours)	**0.8830**	**0.8675**	**0.8830**	**0.8752**

缺陷依据分类实验结果如表 5 所示。在该实验中，BERT 模型基本实现了对分类依据的识别，而其他基准方法则在该复杂多分类问题上表现不佳，难以建模缺陷现象与分类依据之间的关系，其准确率、精确率皆在 0.6 以下。

表 5 缺陷依据分类实验结果

Table 5 Defect classification experiment results

方法	准确率	精确率	召回率	F1 值
TextCNN	0.5805	0.5770	0.5805	0.5581
TextRNN	0.5601	0.5328	0.5601	0.5308
TextRNN-Att	0.5485	0.5228	0.5485	0.5175
TextRCNN	0.5716	0.5720	0.5716	0.5439
FastText	0.5687	0.5639	0.5687	0.5439
DPCNN	0.5409	0.5261	0.5409	0.5093
Transformer	0.5140	0.5196	0.5140	0.4966
BERT(Ours)	**0.6822**	**0.6474**	**0.6822**	**0.6643**

由分类实验结果，可得以下结论：

1) BERT 模型在文本分类任务上具有巨大的潜力，可以有效理解电力领域专业用语、专业知识；

2) BERT 模型更适应少样本环境，能在训练数据有限的情况下捕捉样本特征，且不容易陷入过拟合；

3) BERT 模型更善于处理多样的文本信息与分类场景，更能应对复杂的阅读理解问题。

4　大语言模型的实验结果与分析

4.1 实验数据

在本文模型架构中，大语言模型的主要任务为汇总 BERT 分类信息，针对缺陷现象给出相应的分析与解决方案。该任务原始数据集介绍见第 3 节。本文从中提取了大量主设备缺陷的具体缺陷现象，并用这些现象构建数据集，构成问答数据集。回答的质量将由专家评判获得，评价标准细节见 4.2 节。

4.2 评价指标

目前针对文本生成任务的自动化评估已经取得了显著进展，尤其在文本翻译等任务上，BLEU[22]、ROUGE[23]、BERTScore[24]等方法可以取得与人工评审相近的结果。但在本节的任务背景下，这些方法都难以取得满意的效果，有以下原因：1) 缺陷分析有多个角度，解决方案也往往有多个条目，难以通过简单的重叠度计算或语义匹配进行评价；2) 数据集内针对缺陷的处理方法为人工填写，存在主观性与一致性问题，无法视作有效的答案。为提高评价结果的可信度，本文选用人工评审方法，结合电力问题的特性，综合考虑生成回答的质量及回答在电力问题上的正确度、应用价值，设计了一套以流畅度、一致性、常识逻辑性、回答准确度为主的评价体系，并将满分置为 100 分，其设计如表 6 所示。

表 6 问答任务评分标准

Table 6 Grading criteria for Q&A tasks

评估维度	占比	评估维度	占比
流畅度	0.1	常识逻辑性	0.3
一致性	0.1	回答准确度	0.5

其中，流畅度反映回答的可读性，生成的回答越通顺，越容易阅读，该项得分越高；一致性指回答与电力领域的关联程度，模型的回答应该与电力领域相关，而不应该过多涉及其他领域的知识，否则将在该项内进行扣分；常识逻辑性指回答内容是否符合一般逻辑，有没有容易出现歧义或误会的地方；回答准确度指问题的答案是否正确，如果答案完全错误，则不对其他项进行评分，直接记为 0 分。常识逻辑性和回答准确度的评分细则如表 7 所示。

本文从生成回答中随机选取 200 条，由 3 名电力领域专家分别打分，并取平均值，评定回答的质量，用于评估模型在问答任务上的能力。

4.3 实验结果分析

电力主设备缺陷问答实验结果如表 8 所示。由实验结果可知，只使用大语言模型与增加提示引导相比，在流畅度和一致性上差距不大，但仍高于无提示的情况。在常识逻辑性上，使用了提示的大语言模型比无提示模型提高了 2.75 分，提升了约 12.67%，这说明设计合理的提示可以在客观上提高大语言模型生成结果的表述合理性。而针对回答准确度，有提示模型较无提示模型的得分提高了 2.91 分，提升约 7.77%。该实验结果可以证明提示学习对生成合理的答案有显著的正面影响。

表 7 常识逻辑性和回答准确度的评分细则

Table 7 Scoring criteria for common sense logicality and answer accuracy

评估维度	分数	解释
常识逻辑性	0—10	回答几乎没有逻辑，只是自然语言的堆叠
常识逻辑性	10—20	回答具有一定的逻辑，但可以发现明显的常识性错误或冲突
常识逻辑性	20—30	回答逻辑完备，但可能留存机器生成的痕迹
回答准确度	0—10	回答完全错误，或与问题完全无关
回答准确度	10—20	回答中包含明显的错误、漏洞或对问题的理解有明显偏差
回答准确度	20—30	回答与问题关联不足，但没有明显的错误
回答准确度	30—40	回答与问题强相关，但有细节上的错误，或缺乏可行性
回答准确度	40—50	回答与问题强相关，答案基本正确，且具有一定的可行性

表 8 缺陷问答实验结果

Table 8 Q&A experiment results

评估维度	ChatGLM3(得分)	ChatGLM3 + 提示信息(得分)
流畅度	7.78	7.98
一致性	8.42	8.64
常识逻辑性	21.70	24.45
回答准确度	37.43	40.34
总计	**75.33**	**81.41**

由问答实验，可以得出以下结论：

1) 大语言模型初步具备了一定的电力知识，可以回答简单的电力领域问题，但生成的回答可能有细节错误，或太过宽泛，缺乏实用性；

2) 提示学习有助于提高大语言模型的常识逻辑性和回答准确度，即有助于提高模型的推理能力，帮助模型回答更复杂的问题。

4.4 案例分析

为了进一步说明大语言模型在缺陷问答任务上的良好表现及提示学习的作用，本节选取其中有代表性的案例进行详细阐述。

示例如图 5 所示。该案例提出了一个关于开关

柜温湿度控制器的问题。“控制器无法实现温控功能”未出现在标准库中，属于罕见故障，考虑到开关柜升温是一个长期过程，且不会立即影响正常供电和从业人员的人身安全，可定为一般缺陷。

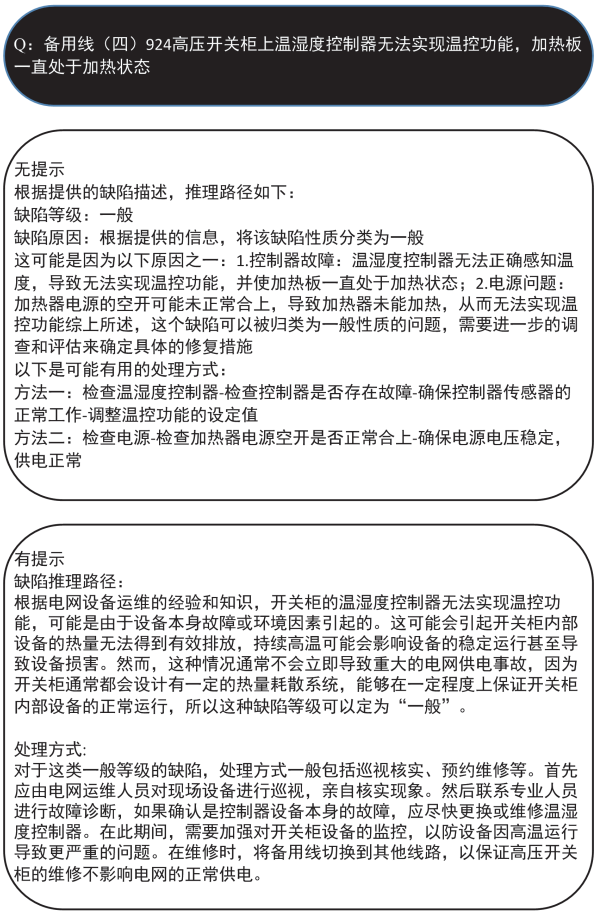

图 5 示例 1

Fig. 5 Case analysis 1

根据回答内容可知，无论是否使用提示，大语言模型都可以正确判断缺陷等级。但在不使用提示的情况下，模型对“推理路径”的认识有偏差，只分析了缺陷原因，没有一步步说明该原因导致的后果，无法形成逻辑链路。同时，对“无法实现温控功能”的认识过于广泛，从过冷和过热两个角度分析，这明显是冗余且没有必要的。而在增加了提示的情况下，大语言模型可以简要分析缺陷原因，并针对缺陷等级分类结果给出合理的推理路径。

综上，使用提示学习方法有助于大语言模型更深入理解文本语义，从而提高回答的质量，使模型能更好地完成“主设备缺陷诊断”这一核心任务。针对问答结果而言，该方法可有效提高大语言模型的推理能力，增加结果的逻辑性、准确度和有效性。

4.5 大语言模型的其他实验

提示学习允许大语言模型在不重新训练的情况下回答特定领域问题、引用外部数据信息、执行多种下游任务。本文尝试从两个视角进一步挖掘大语言模型在电力领域的应用潜力。首先，大语言模型具有卓越的文本生成能力，可以用于数据集的生成、增强；其次，大语言模型本身具备语义理解能力与电力领域知识，可以在不进行参数微调的情况下回答电力领域问题，即零样本学习。

4.5.1 大语言模型用于缺陷文本增强

在电力故障记录单中，缺陷现象多为人工描述，存在大量缩写、简写情况，不便直接用于统计或模型训练。而大语言模型有赖于丰富的预训练资源，可以理解多种格式的数据，并将这些数据整理为统一的表格或自然语言，直接方便系统整合与从业人员阅读，且有望为模型训练提供更高质量的数据。

为有效利用大语言模型进行缺陷文本增强，本文构建了一种背景、原则、模板、语言风格四位一体的大语言模型提示模板框架，如图 6 所示。在该提示模板的作用下，大语言模型对“硅胶变色”缺陷进行了更详细的说明，将缩写的“主变”复原为“主变压器”，并简要说明了该现象可能导致“吸湿性能下降”，因此可能需要对硅胶进行及时更换，与原始文本相比，包含了更多的关键信息。

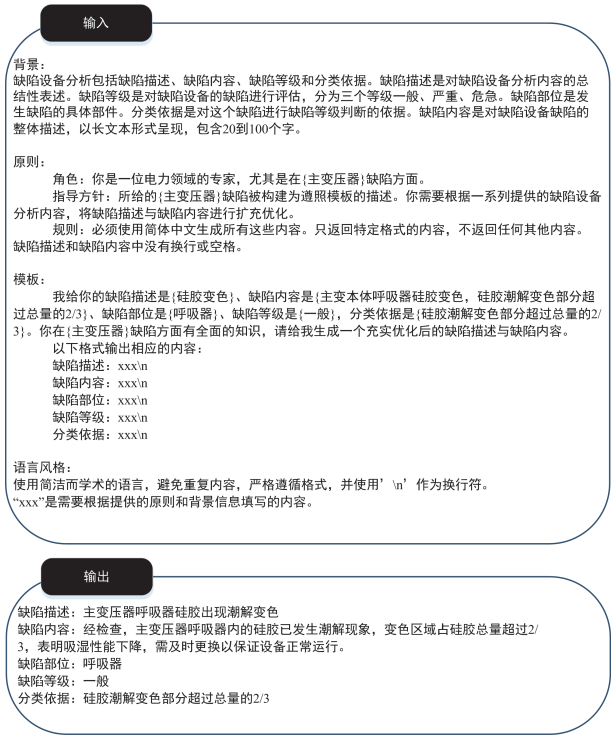

图 6 缺陷文本增强提示模板

Fig. 6 Prompt template for defect data augmentation

4.5.2 基于大语言模型的缺陷等级分类

基于大语言模型提示的电力问答实验结果已证明，大语言模型本身具备了一定的电力知识与领域专业词汇理解能力。本节尝试只通过添加提示的方法，在不训练模型的前提下进行缺陷等级分类。

该方法以提示工程为核心只使用原始的角色扮演提示和思维链提示，如，发生缺陷：“#2 主变 220 kV 侧门型架上有一鸟窝”，可将这个现象填写入提示模板中，查询缺陷的等级和分类依据，通过大语言模型问答获得结果。该案例的实践如图 7 所示。

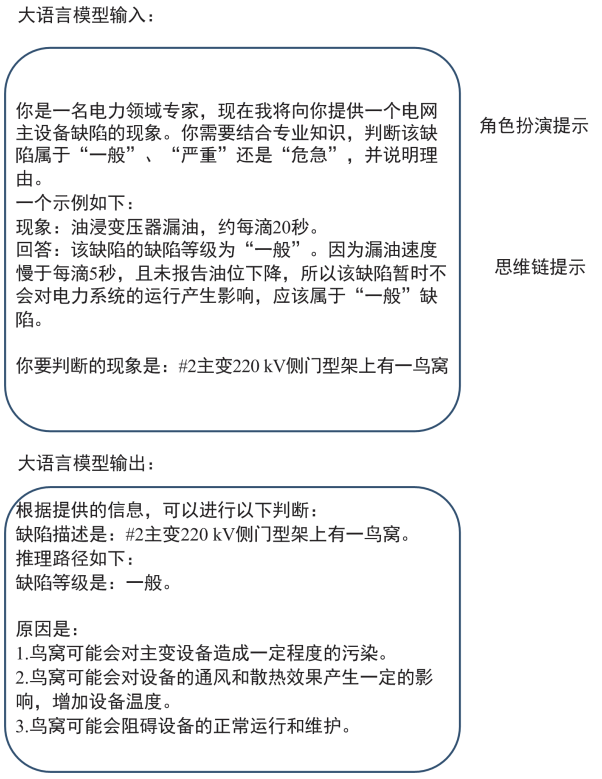

图 7 大语言模型用于缺陷等级分类

Fig. 7 LLM for defect levels classification

“出现鸟窝”并非电力系统主设备的典型缺陷，但根据“一般”缺陷的定义，一般缺陷指该缺陷暂未表现出危及人身安全或电力系统运行的趋势。该缺陷的分级应该为“一般”，大语言模型给出了正确的答案，且针对这一分类结果给出了合理的解释。

5 结论

针对传统主设备缺陷诊断依赖人工、自动化程度低、效率低下等问题，本文提出了一种基于 BERT 等语言模型的两阶段诊断框架。第一阶段，通过微调 BERT 模型执行分类任务，初步理解缺陷；第二阶段，利用提示工程，激发大语言模型的推理潜能，使之生成高质量的缺陷分析和处理意见。模型不但在分类任务上取得了卓越的成果，而且能针对电力主设备缺陷问题生成可靠可行的答案。之后，本文又从文本增强和零样本学习两个角度探究了大语言模型在电力领域的其他应用。

然而，大语言模型本身仍存在内部领域知识匮乏、答案过于宽泛的问题，如何通过参数高效微调获得一个电力大语言模型，是下一步研究的重点。

参考文献

[1] SHChERBATOV I, LISIN E, ROGALEV A, et al. Power equipment defects prediction based on the joint solution of classification and regression problems using machine learning methods[J]. Electronics, 2021, 24(10).

[2] 李元, 李睿, 林金山, 等. 基于字词混用集成模型的电力变压器缺陷记录文本挖掘方法[J]. 电力工程技术, 2024, 43(6): 153-162.
LI Yuan, LI Rui, LIN Jinshan, et al. Character-word level ensemble integrated model for power transformer defect recording text mining method[J]. Electric Power Engineering Technology, 2024, 43(6): 153-162.

[3] GAO Q, ZHONG C, WANG Y, et al. Defect analysis of the same batch of substation equipment based on big data analysis algorithm[C] // IOP Conference Series: Earth and Environmental Science, November 14-15, 2020, Shenyang, China.

[4] 陈岳, 陈翔宇, 邓军, 等. 基于缺陷记录的直流主设备主要运行缺陷统计与分析[J]. 高压电器, 2015, 51(8): 180-185.
CHEN Yue, CHEN Xiangyu, DENG Jun, et al. Statistical analysis of basic defects in conventional DC main equipment based on defect records[J]. High Voltage Apparatus, 2015, 51(8): 180-185.

[5] 朱保军, 咸日常, 范慧芳, 等. WRSR 与改进朴素贝叶斯融合的变压器故障诊断技术研究[J]. 电力系统保护与控制, 2021, 49(20): 120-128.
ZHU Baojun, XIAN Richang, FAN Huifang, et al. Transformer fault diagnosis technology based on the fusion of WRSR and improved naive Bayes[J]. Power System Protection and Control, 2021, 49(20): 120-128.

[6] 欧阳鑫, 李志斌. 基于样本扩充和特征优选的 IGWO 优化 SVM 的变压器故障诊断技术[J]. 电力系统保护与控制, 2023, 51(18): 11-20.
OUYANG Xin, LI Zhibin. Transformer fault diagnosis technology based on sample expansion and feature selection and SVM optimized by IGWO[J]. Power System Protection and Control, 2023, 51(18): 11-20.

[7] 俞伊丽, 张展耀, 接晓霞, 等. 基于知识图谱与 SCD

文件的智能变电站二次检修安全措施自动生成技术研究[J]. 电力系统保护与控制, 2024, 52(2): 129-142.
YU Yili, ZHANG Zhanyao, JIE Xiaoxia, et al. Automatic generation technology of secondary safety measures in an intelligent substation based on a knowledge graph and SCD files[J]. Power System Protection and Control, 2024, 52(2): 129-142.

[8] ZHU J, LI G, GUO Y, et al. Real-time risk-averse dispatch of an integrated electricity and natural gas system via conditional value-at-risk-based lookup-table approximate dynamic programming[J]. Protection and Control of Modern Power Systems, 2024, 9(2): 47-60.

[9] 余建明, 刘赫, 单连飞, 等. 基于ALBERT和RE2融合模型的电网调度意图识别方法[J]. 电力系统保护与控制, 2022, 50(12): 144-151.
YU Jianming, LIU He, SHAN Lianfei, et al. Method of power grid dispatch intention recognition based on ALBERT and RE2 fusion model[J]. Power System Protection and Control, 2022, 50(12): 144-151.

[10] 晏鹏, 黄晓旭, 黄玉辉, 等. 基于 BERT-DSA-CNN 和知识库的电网调控在线告警识别[J]. 电力系统保护与控制, 2022, 50(4): 129-136.
YAN Peng, HUANG Xiaoxu, HUANG Yuhui, et al. Online alarm recognition of power grid dispatching based on BERT-DSA-CNN and a knowledge base[J]. Power System Protection and Control, 2022, 50(4): 129-136.

[11] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arxiv preprint arxiv:1810.04805, 2018.

[12] 田波, 张越, 蒙飞, 等. 电网故障处置信息自适应理解框架及关键技术[J]. 中国电力, 2024, 57(7): 188-195.
TIAN Bo, ZHANG Yue, MENG Fei, et al. Adaptive understanding framework and key technology of power grid fault disposal information[J]. Electric Power, 2024, 57(7): 188-195.

[13] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014: 1532-1543.

[14] LILLEBERG J, ZHU Y, ZHANG Y. Support vector machines and word2vec for text classification with semantic features[C] // 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), July 6-8, 2015, Beijing, China: 136-140.

[15] 张金营, 王哲峰, 谢华, 等. 基于知识图谱与大语言模型的电力行业知识检索分析系统研发与应用[J]. 中国电力, 2024, 57(12): 198-205.
ZHANG Jinying, WANG Zhefeng, XIE Hua, et al. Development and application of a knowledge retrieval and analysis system for the power industry based on knowledge graph and large language model[J]. Electric Power, 2024, 57(12): 198-205.

[16] ZHANG Z, ZHANG A, LI M, et al. Automatic chain of thought prompting in large language models[J]. arxiv preprint arxiv:2210.03493, 2022.

[17] JOULIN A, GRAVE E, BOJANOWSKI P, et al. Fasttext.zip: compressing text classification models[J]. arxiv preprint arxiv:1612.03651, 2016.

[18] ZHANG M, PANG J, CAI J, et al. DPCNN-based models for text classification[C] // 2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud) / 2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom), July 1-3, 2023, Xiangtan, China: 363-368.

[19] CHEN Y. Convolutional neural network for sentence classification[D]. Ontario, Canada: University of Waterloo, 2015.

[20] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2015, 29(1).

[21] TEAM G L M, ZENG A, XU B, et al. Chatglm: a family of large language models from GLM-130b to GLM-4 all tools[J]. arXiv e-prints, 2024.

[22] 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, July 2002, Philadelphia, USA: 311-318.

[23] LIN C, OCH F J. Rouge: a package for automatic evaluation of summaries[C] // Text Summarization Branches Out, California, USA: 74-81.

[24] ZHANG T, KISHORE V, WU F, et al. Bertscore: evaluating text generation with Bert[J]. arXiv Preprint arXiv: 1904.09675, 2019.

收稿日期：2024-04-21； 修回日期：2024-09-23

作者简介：

杨 虹(1982—)，女，通信作者，硕士，高级工程师，研究方向为输变电设备运维技术；E-mail: 199132003@qq.com

孟晓凯(1987—)，男，博士，高级工程师，研究方向为高压电缆绝缘状态检测技术；E-mail: 2381321367@ qq.com

俞 华(1980—)，男，硕士，正高级工程师，研究方向为输变电设备故障诊断技术。E-mail: 3871552206 @ qq.com

(编辑 张 颖)