文章编号:1671-4598(2026)10-0008-08 DOI:10.16526/j.cnki.11-4762/tp.2026.10.002 中图分类号:TP206+.3 文献标识码:A

基于大语言模型的潜油电泵异常井智能诊断方法

李泓臻
(中海油能源发展股份有限公司 工程技术分公司,天津 300452)

摘要:针对潜油电泵(ESP)在复杂工况和油井环境下,早期异常特征微弱、井间数据分布漂移显著与单一模型容易误检漏检等问题,提出一种基于大语言模型的ESP异常检测方法;结合多种工况数据与对应油井背景知识,构建上下文输入;建立可扩展的ESP领域指令微调数据集,基于多种中小参数开源模型,通过LoRA指令微调构建差异化专家ESP异常诊断模型集合,并由单独的仲裁系统对专家输出集合进行链式推理与一致性校验,生成最终判别结果;以某钻井平台ESP数据为对象进行了实验研究,验证了所提出方法的可行性与稳定性,准确率远优于其他主流的机器学习或深度学习方法。

关键词:潜油电泵(ESP);异常检测;大语言模型;LoRA微调;链式推理(CoT);上下文增强

Intelligent Diagnosis Method for Abnormal Wells of ESP Based on LLMs

LI Hongzhen
(CNOOC Energy Technology & Services Limited Engineering Technology Branch, Tianjin 300452, China)

Abstract: To address the issues of electric submersible pumps (ESP) operating under complex working conditions and oilwell environments, including weak early abnormal features, significant inter-well data distribution drift, and the tendency of a single model to produce false positives and false negatives, an ESP anomaly detection method based on large language models (LLMs) is proposed. Multiple operating condition data are integrated with corresponding oil well background knowledge to construct contextual inputs, with an extensible ESP-domain instruction fine-tuning dataset built. Besides, based on multiple small to medium-parameter open-source models, a differentiated expert ESP anomaly diagnosis set is established by low rank adaptation (LoRA) -based instruction fine-tuning, and a separate arbitration system performs chain-of-thought (CoT) reasoning and consistency verification on the expert output set to generate the final judgment decision. Experimental studies are conducted on the ESP data from a certain drilling platform, validating the feasibility and stability of the proposed method, and this method is significantly superior in the accuracy to other mainstream machine learning or deep learning methods.

Keywords: ESP; anomaly detection; LLMs; LoRA fine-tuning; CoT; context augmentation

0 引言

作为油田人工举升系统的核心装备,潜油电泵(ESP, electric submersible pump)的运行状态直接关系到油气产量与生产能耗。由于长期工作在井下高温、高压、多相流耦合的复杂环境中,易发生气锁、机械磨损、供液不足等多种故障。且这些异常具有早期信号弱、变量耦合强、跨井差异大等[1]特点,使得实现精准、及时且可解释的智能诊断成为油田运维领域长期面临的挑战[2]。现有的ESP故障诊断方法[3]多依赖于物理机理模型、专家规则库或基于浅层统计特征的机器学习[4]模型。此外,ESP工况数据变量复杂,时间跨度较高,单变量[5]或局部时序模型[6]难以构建完整的因果证据链,诊断结果的可解释性与工程可信度往往不足。大语言模型(LLMs, large language models)凭借其卓越的

收稿日期:2026-02-02; **修回日期**:2026-03-18。
基金项目:中海油能源发展科技重大专项(HFKJ-ZD-GJ-2024-01-02)。
作者简介:李泓臻(1989-),男,大学本科,工程师。
引用格式:李泓臻. 基于大语言模型的潜油电泵异常井智能诊断方法[J]. 计算机测量与控制,2026,34(10):8-15,24.

跨模态语义理解与逻辑推理能力，为复杂场景下的工业诊断[7]提供了全新的技术路径。然而，将通用 LLMs 直接用于 ESP 场景仍面临两类关键障碍：1）异常检测高度依赖油藏地质、井史、设备配置等领域先验知识，而通用 LLMs 缺乏对此类上下文信息的具体理解；2）在复杂多变的工况与油井环境差异下，单一模型生成的异常检测结果稳定性较差，不确定性高。因此，如何设计适合 ESP 异常检测场景的上下文输入，并通过系统结构设计提升检测的稳定性，是推动 LLMs 在 ESP 运维中落地的关键。本文的主要贡献包括：探索性地将大语言模型引入 ESP 异常检测与类型识别任务，并提出面向 ESP 场景的多模态语义融合思路；构建多专家－仲裁式协同大语言模型诊断框架，通过模型集成与推理仲裁有效缓解单模型框架的偏差，从而增强决策稳定性与可解释性；建立可扩展的 ESP 领域指令微调数据集，形成支持后续模型迭代与扩展的数据基础。

1　面向 ESP 异常检测的多专家－仲裁式系统

1.1　ESP 工况数据与背景信息准备

在潜油电泵井的智能诊断场景中，本文使用的数据由两大模态构成：1）井口与井下的时序数值工况数据；2）油井的背景信息，如油藏地质、井史、设备配置等，具体如表 1 所示。

表 1　ESP 异常检测数据的模态组成及变量说明

类别	变量
工况信息	CSG_PRESS 套管压力
	PUMP_INLET_PRESS 泵入口压力(吸入压力)
	PUMP_OUTLET_PRESS 泵出口压力
	PUMP_OUT_IN_PRESS 泵进出口压差
	PROD_OUT_IN_PRESS 生产系统进出口压差
	FLOW_LINE_PRESS 流程管线压力
	WH_TEMP 井口温度
	OIL_PRESS 油压
	PUMP_INLET_TEMP 泵入口温度
	PUMP_CURRENT 泵电机电流
	LEAKAGE_CURRENT 漏电电流/泄漏电流
	MOTOR_POWER 电机功率
	MOTOR_TEMP 电机温度
	PUMP_TORQUE_CURRENT 泵扭矩电流
	VIBRATION 振动
	PUMP_VOLTAGE 泵电机电压
	NOZZLE_DIAMETER 喷嘴/节流嘴直径
	PUMP_FREQ 泵频率(变频器频率)
背景信息	井型、完井方式、地层压力、原油黏度、含砂情况
	完井方式、投产时间、历次修井、调参记录、历史故障类型与处理情况
	地层结构、渗透率、孔隙度、原始地层压力、油藏类型、含砂情况
	泵型号、级数、设计扬程、安装深度、电机功率

ESP 工况原始数据来源于某钻井平台系统内部实时数据服务接口，其返回结构为以井号为键的字典，每口井对应一组工况时序记录。本研究设定采样间隔 3 600 s。考虑到接口在采样终止时刻返回的数据可能存在不完整或实时更新的情况，预处理过程中剔除了每个时间序列最后一个时间点的记录，并重置了索引。通过上述处理，确保所有井的数据在全局时间轴上实现严格对齐。

在完成等间隔重采样后，对时间信息进行标准化处理，包括统一转换至标准时区、剔除重复时间戳与乱序记录。针对明显不符合物理规律的异常值，如负压力、负功率及极端振动峰值等，本文采用截断处理或将其标记为缺失值，并对所有缺失数据应用线性插值进行填补。随后，将清洗后的工况时序数据按井号分组，得到针对每口井 i 的统一的等间隔工况序列 $X^{(i)}$。

为克服大语言模型在处理长时序数据时存在效率低下与鲁棒性不足的问题，同时为异常检测提供必要的局部上下文信息，本文引入滑动窗口构造样本。具体而言，对于每口井 i，在时间点 t 处定义一个长度为 L 的历史窗口：

$$X_t^{(i)} = x_t^{(i)}, \cdots, x_{t+L-1}^{(i)} \tag{1}$$

其中：$x_t^{(i)}$ 为 d 维的 ESP 测量与控制量，具体特征如表 1 所示。

为消除不同井之间量纲与量级的差异，在窗口构造后，本文对数值特征进行标准化处理。对于压力与电流等常规物理量，采用按井标准化的方式，即针对每口井 i 单独计算其均值与标准差，再进行 Z-score 归一化[8]。对于如振动、泄漏电流等呈长尾分布的特征，则先进行对数变换以压缩动态范围，再执行上述标准化。

记油井的背景信息为 $B^{(i)}$，包含井的基本信息、油藏地质资料、井史事件以及设备配置等长期先验知识。与工况序列数据不同，$B^{(i)}$ 在给定井内随时间基本保持不变，并且为纯文本数据，用于进行 ESP 状态的辅助判断，例如，渗透率中低且易出砂的油藏更容易在高压差工况下出现泵吸入口堵塞或管柱磨损；安装深度与油藏压力共同影响气锁和气体侵入的风险。

1.2　多模态数据融合与问题定义

在完成两类模态整理后，本文采用模板化方式构造统一的诊断上下文，从而完成动态工况数据与静态背景的拼接与融合。对每个样本，输入由任务指令、工况窗口描述与油井背景描述三部分组成。统一输入可表示为：

$$C_t^{(i)} = S^{\text{candidate}}, X_t^{(i)}, B^{(i)} \tag{2}$$

其中：任务指令 $S^{\text{candidate}}$ 用于约束中小专家模型输出异常判别结果及原因解释；工况窗口 $X_t^{(i)}$ 用于呈现当前时段的动态运行特征；油井背景描述 $B^{(i)}$ 用于提供跨时

间稳定的工程先验。通过将两类模态拼接于同一诊断上下文中，模型能够对异常现象进行条件化分析，从而减少仅依赖工况信号时可能出现的误判。

设 $M=[A, B, C, \cdots]$ 表示用于 ESP 异常检测任务的专家中小规模 LLM 集合。给定任务输入 $C_t^{(i)}$，每个单独的专家模型产生的结构化异常检测响应为：

$$y_t^{(i),m} = f^m(C_t^{(i)}) = [\hat{y}_t^{(i),m_{abn}}, e_t^{(i),m}], m \in M \quad (3)$$

其中：$\hat{y}_t^{(i),m_{abn}} \in \{0,1\}$，表示异常状态判别变量，0 表示状态正常，1 表示状态异常；$e_t^{(i),m}$ 为模型对可能出现的异常类型的分析，以及做出异常判断的原因解释。

为整合多个候选检测模型的输出，并克服单模型偏差以提升决策的一致性与可解释性，本文引入一个专门的仲裁系统（LLM-as-a-Judge）J，与前述执行异常检测的模型不同，仲裁模型 J 的输入不仅包括原始任务输入 $X_t^{(i)}$ 与 $B^{(i)}$，新的任务指令 $S^{\text{candidate}}$，还包含先前专家中小语言模型集合的输出 $y_t^{(i),m}$。结合三者的输入，仲裁模型给出最终的异常检测判断结果 $y_t^{(i),J}$。

本文的目标是构建一组主力 ESP 异常检测任务模型 $\{f_m\}$ 与仲裁系统 J，使得最终的异常检测输出结果 $y_t^{(i),J}$ 符合实际，与人类专家判断尽可能一致。

1.3 ESP 异常检测专家模型集合

为提升异常检测响应的多样性与鲁棒性，本文构建了一个由多个中小规模语言模型组成的专家集合，所有模型均基于 ESP 异常检测数据集进行了微调。选择中小规模模型主要基于两方面考虑：一方面，该规模模型已具备良好的通用语言理解、指令跟随与结构化输出能力，适用于基于工况时序描述与背景知识的异常分析任务；另一方面，其规模适中，便于后续通过监督微调进一步融入专家知识。

为确保模型在决策层面的互补性，本文有意识地选取了来自不同技术路线的开源基座模型。尽管这些模型在参数规模上相近，但在预训练语料构成、推理风格和响应倾向上存在显著差异。在 ESP 异常检测任务中，这种差异并不体现为性能上的优劣之分，而更多表现为对同一输入的关注重点不同。例如，有的模型更倾向于根据局部工况变化作出快速判断，有的更注重挖掘背景条件与异常机理之间的因果关联。正是这种源于异构性的多样性，为后续通过多专家仲裁机制形成互补判断、提升整体决策质量提供了重要基础。

在专家集合构建方式上，本文将各个中小语言模型视为平级专家。这些专家模型的响应共同构成了一个丰富的证据池。

后续仲裁模型 J 的思维链[9]（CoT，chain-of-thought）将利用多专家模型响应，分析其中异同，从而获得更丰富的上下文信息。相较于传统的单模型方法，本方法的优势在于能够通过预先提供的多样化证据，更自然与深入[10]地引导模型覆盖更多样的推理空间，减轻模型的幻觉和偏见。

如图 1 所示，3 个候选模型对同一输入给出了基于不同证据维度与思考的差异化评判与解释。模型 A 识别到泵吸入口压力在短时间窗内呈现持续上升趋势，并考虑油井的产量信息，并在同一异常现象的基础上，给出多种异常原因的猜想。而模型 B 则建立了工况信号与地质信息的联系，结合地层隔夹层、供液环境等信息，对潜在气体扰动进行推断。模型 C 在同一异常现象下，结合含砂量的背景信息，认为“泵磨损”的异常可能性更高。尽管归因路径不同，三者均一致认为该样本呈现异常迹象。因此，在后续的思维链推理中，仲裁模型 J 将分析各个专家模型输出的相同和不同之处，并将其作为思维链下一步的输入，为下一步的最终决策提供了更丰富与全面的上下文内容。

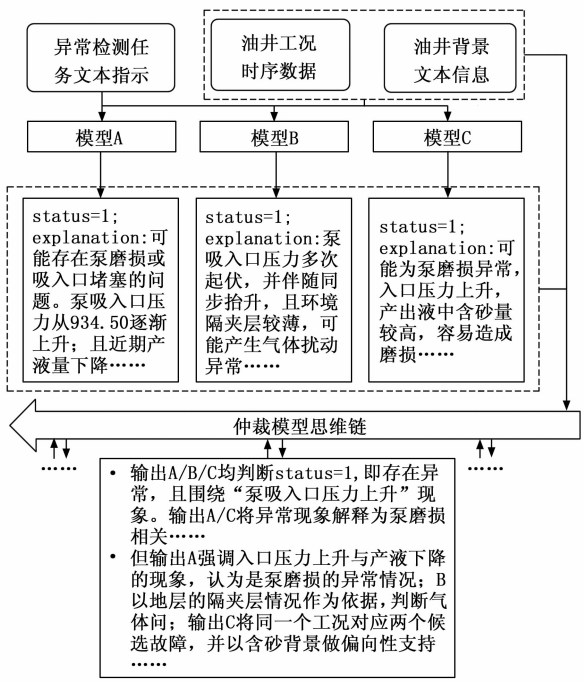

图 1 多专家候选模型差异化诊断输出示意

利用多个 ESP 异常检测专家模型，生成多样化的响应集合的核心优势在于，它系统性地利用了不同模型在信号敏感维度与归因逻辑上的差异性。这种差异促使生成的初步判断能够从多维的工况分析空间出发，为仲裁阶段的链式推理提供了丰富且互补的证据集合。通过对这些证据进行交叉验证与协同分析，框架能够有效纠正单一模型的偏见，从而显著提升对边界工况与弱异常信号的辨识鲁棒性。从机制上看，这一过程不仅降低了漏检与误检的概率，也因保留了多元推理路径，而极大地增强了最终诊断结论的可解释性与决

策的可追溯性。

1.4　ESP 异常检测专家模型的微调

为提升各个中小规模 LLM 在 ESP 异常检测任务中的性能，本文构建了一个基于现场真实工况与油井背景知识的监督式指令数据集，并采用参数高效微调方法（LoRA）对其进行微调。

在监督式指令微调数据集中，每条用于训练的对话样本 $D_t^{(i)}$ 由两部分组成：

$$D_t^{(i)} = [C_t^{(i)}, y_t^{(i),gt}] \tag{4}$$

其中：$C_t^{(i)}$ 为训练样本中输入给模型的上下文内容，即由任务指令、井背景信息、工况时序组成的对话输入。$y_t^{(i),gt}$ 为监督信号，即训练样本对应的真实标注输出，即人工专家的标准诊断回答。

ESP 异常检测微调数据集的对话样本示例如下所示：

```
{
  "index": 6,
  "instruction": "你是潜油电泵 ESP 智能诊断助手。请基于设备工况时序数据与静态背景信息,对设备状态进行诊断,并输出结构化结果……",
  "input": {
    "work_condition_series": {
    "pressure_signals": "<pressure_time_series>",
    "temperature_signals": "<temperature_time_series>",
    "electrical_signals": "<current_voltage_power_time_series>",
    "mechanical_signals": "<torque_vibration_time_series>",
    "control_signals": "<frequency_time_series>",
    "structural_parameter": {
  "nozzle_diameter": "<value>"
    }
  },
  "static_background_info": {
    "basic_info": "<basic_well_information>",
    "reservoir_geology": "<reservoir_and_wellbore_information>",
    "well_history": "<historical_operation_and_fault_records>",
    "equipment_config": "<pump_and_motor_configuration>"
  }
    },
  "output": {
  "status": "<0_or_1>",
  "explanation": "<diagnosis_explanation>"
  }
}
```

在监督式指令微调阶段，将拼接后的对话序列输入中小规模 LLM，并以参考输出作为监督信号，最小化目标响应 token 的交叉熵损失：

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{(i,t)\in Z}\sum_{\tau=1}^{|y_t^{(i)}|}\log p_\theta[y_{t,\tau}^{(i)} \mid x_t^{(i)}, y_{t,<\tau}^{(i)}] \tag{5}$$

其中：Z 表示所有油井在不同时间序列下的样本集合，τ 为 token 的位置索引，$|y_t^{(i)}|$ 表示响应序列的 token 长度。

为在有限标注成本与可控训练开销下，提升模型对 ESP 异常机理的适应能力，本文采用 LoRA 方法，将通用基座模型迁移至目标领域。具体而言，LoRA 不直接更新原始权重矩阵 $\boldsymbol{W}_0$，而是在冻结基座模型参数的基础上，引入可训练的低秩增量矩阵 $\Delta\boldsymbol{W}$。训练阶段仅更新该增量矩阵，而保持基座模型参数冻结；推理阶段则可将其合并回原始权重。该机制使得模型在保留通用语言知识的同时，能够以较低的参数开销学习 ESP 异常检测任务所需的领域知识。

在具体实现上，本文仅将 LoRA 适配层插入注意力模块中的 Query 投影矩阵 $\boldsymbol{W}_Q$ 与 Value 投影矩阵 $\boldsymbol{W}_V$。由于 $\boldsymbol{W}_Q$ 与 $\boldsymbol{W}_V$ 直接参与注意力分配与信息聚合过程，对模型领域适应能力具有更高敏感性，因此在该位置引入低秩适配可在训练显存开销较低、参数效率较优的前提下，取得良好的微调效果。

以 Qwen2.5-7B 为例，微调前的输出示例如下：

“output”：{

“status”：1.

“explmnation”：“最近一段时间内泵的状态不太稳定，电流、振动和温度这些指标有上升的趋势。结合井史里之前出现过扭矩偏高的情况，这次可能也是类似的机械问题，整体判断更偏白异常。”

微调后输出示例如下所示：

“output”：{

“status”：1.

“explmnation”：“故障前 24 小时内，出口压差整体保持相对平稳，未见与之对应的明显水力工况改善或恶化，但电机扭矩电流与振动幅值呈现持续上行趋势，电机温度同步抬升，显示负载主要来自轴系/机械侧而非纯水力侧。结合该井并史曾出现扭矩偏高并采取减载处理，以及井筒附近存在一定并斜和狗腿度，含砂风险中等等背景，轴系长期受力与磨蚀条件客观存在，该“水力变化有限＋机械负荷与振动升高”的组合模式与轴系疲劳、裂纹扩展直至轴断的演化特征具有极强的一致性。”

1.5　ESP 异常检测仲裁系统

在获得各中小规模专家 LLM 输出 $y_t^{(i),m}$ 后，关键

环节在于对这些异常检测结果进行有效分析与综合，以支撑最终决策。由于 ESP 异常涉及多工况变量耦合关系，且受井间地质条件与设备差异的显著影响，各专家模型难免存在错检漏检等问题。若直接采用随机选取或简单多数表决，容易引入不稳定的主观偏好，难以得到最优结果。因此，本文通过基于大语言模型的链式推理框架对专家模型输出集合进行归纳与分析，建立仲裁系统 J。ESP 异常检测仲裁系统指令数据集的数据样本框架如下所示：

```
{
    "role_definition": {
    "name": "ESP 异常诊断仲裁专家……",
    "description": "<TASK_DESCRIPTION>"
},
    "context_inputs": {
    "basic_background": "<STATIC_BACKGROUND_INFO_JSON>",
    "timeseries_window": "<WORK_CONDITION_SERIES_JSON>"
},
    "expert_responses": [
    "<EXPERT_A_RESPONSE>",
    "<EXPERT_B_RESPONSE>",
    "<EXPERT_C_RESPONSE>",
    "..."
    ],
    "reasoning_workflow": [
    "要点提炼",
    "一致性校验",
    "跨模型对比与关联分析",
    "归并共识与解释生成"
],
    "output_requirements": {
    "format": "json",
    "status": "0 | 1",
    "explanation": {
    "共识提炼": "<SHARED_FINDINGS>",
    "分歧对比": "<DIFFERENCE_ANALYSIS>",
   "最终诊断与依据": "<FINAL_DECISION>"
    }
},
    "generation_instruction": "<FOLLOW_WORKFLOW_AND_RETURN_JSON_ONLY>"
}
```

为有效整合各个专家模型输出的评判信息，本文选用性能更强的 DeepSeek-V3[11] 作为仲裁系统 J 的基座模型。为了使 J 在复杂推理中保持逻辑连贯与步骤透明，本文采用提示词工程（Prompt Engineering）引导[12]模型遵循 CoT 推理模式。如图 2 所示，该模式要求模型将推理过程分解为多个中间步骤，从而构建更为可靠的群体证据语境，并作出最终评判。

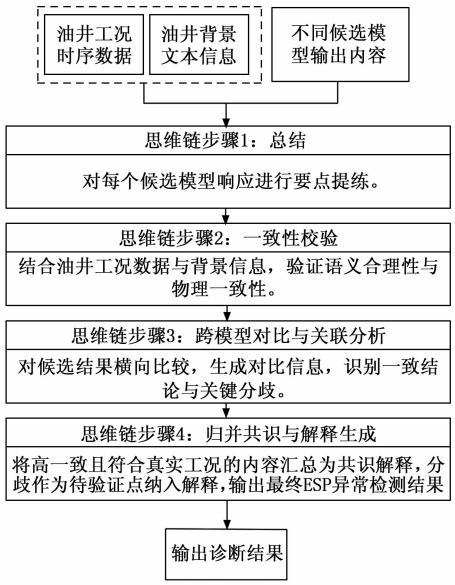

图 2 仲裁系统的构建与推理流程示意图

具体而言，对于由多个中小规模 LoRA 专家模型构成的集合 M，将在相同输入下生成异常检测结果候选输出集合 $\{y_t^{(i),m}\}_{m\in M}$，仲裁系统 J 的输入构成定义为：

$$\widetilde{C}_t^{(i)} = [S^J, X_t^{(i)}, B^{(i)}, \{y_t^{(i),m}\}_{m\in M}] \tag{6}$$

在思维链推理约束下，仲裁系统 J 首先将总结不同专家模型生成的响应内容，再基于输入的油井工况数据与背景信息，对 $\{y_t^{(i),m}\}_{m\in M}$ 是否符合实际数据与背景进行校验。随后，J 将对候选输出进行横向比较，为每个输出 $y_t^{(i),m}$ 生成与其他输出的对比信息，以识别一致结论与关键分歧。最后，仲裁系统 J 将一致性较高，且符合真实工况与油井背景信息的部分归并为群体共识的解释内容，并将仍存在分歧的部分归纳为待验证内容，陈列在解释 $e_t^{(i),J}$ 中。基于上述内容，给出最终的 ESP 异常检测结果与最终解释。

2 实验与结果分析

2.1 实施细节

为提升异常检测结果的多样性与鲁棒性，本文基于大语言模型构建 ESP 异常检测系统。在模型选型上，选取多个开源的中等规模基座模型进行微调，包括 Qwen2.5-7B[13]、Mistral-7B[14]、Llama-3.1-8B[15]、GLM-4-9B-Chat[16] 与 InternLM2.5-7B[17]，用于生成异常检测结果集合。

各模型的选择基于其在通用能力与任务适配性方面的差异化优势：Qwen2.5-7B 在中英文指令理解与工程

微调适配方面表现均衡，适合作为工业诊断任务的稳健基线；Mistral-7B 具备较强的压缩建模能力与较高推理效率，可为多专家集合提供区别于中文主导模型的响应风格；Llama-3.1-8B 作为通用能力较强的开源模型，在指令跟随与上下文理解具有良好基础，能够为异常原因分析提供稳定的语言能力支撑；GLM-4-9B-Chat 在对话式指令响应与复杂任务表达方面具有一定优势，有助于提升诊断结果的完整性与可读性；InternLM2.5-7B 则提供了另一条中文与多语言开源模型路线，增强了候选集合在模型来源上的多样性，从而便于考察不同预训练体系与对齐方式对 ESP 异常检测任务的影响。

训练与推理在 4 张 NVIDIA A800 GPU 上完成，操作系统为 Ubuntu 20.04.6 LTS，CUDA 版本为 12.6。本文基于 PyTorch 与 HuggingFaceTransformers 框架实现 LoRA[18]。所有中小专家模型均采用相同的监督式指令微调流程。本文采用一组常见且稳定的 LoRA 超参数组合，秩 r 设置 8，缩放系数 α 为 16，并使用轻量 dropout，参数取 0.1，以缓解过拟合风险。优化器为 AdamW，学习率为 1×10^{-4}，batch size 为 32，epoch 数量为 3，weight decay 为 0.01，warm-up ratio 为 0.03，最大序列长度设为 2 048。

在数据划分方面，为评估模型的跨油井泛化能力并避免同一口油井内的信息泄漏，本文首先对每口油井的数据按 8∶1∶1 的比例划分训练集、验证集与测试集，随后再进行跨油井组合。该划分确保同一口井的不同时间窗口不会同时出现在训练集与评估集中。

在推理阶段，本文额外调用 DeepSeek-V3.2 的“思考模式”（Thinking Mode）对模型输出进行后处理，旨在纠正可能存在的字段缺失或格式偏移问题，从而避免将纯格式差异误判为模型能力差异。

2.2　实验测试

本文对 ESP 异常检测任务的评估采用异常二分类的评价方式，主要汇报异常识别的精确率（Precision）、召回率（Recall）与 F_1 共 3 项。

为评估 ESP 异常检测任务中引入油井背景信息所带来的性能增益，并确保与基线模型在相同输入条件下进行公平对比，实验对于上述 5 种小参数 LLM，在微调阶段，设置两种输入与标注配置：其一为仅包含 ESP 工况时序数据的基础输入，相应数据集表示为：

$$D_t^{s,(i)} = (S^{s,\text{candidate}}, X_t^{(i)}, y_t^{(i),s,gt}) \tag{7}$$

其中：监督响应 $y_t^{(i),s,gt}$ 已剔除与油井背景知识相关的内容。该配置旨在考察中小参数 LLM 仅依赖短期工况信号完成异常识别的基本能力，方便与基线文献对比。其二为完整的多模态输入数据 $D_t^{(i)}$，即在工况时序数据基础上同时引入油井背景信息，用于评估背景知识对检测性能的提升效果。

上述数据集设置便于量化观测引入油井背景信息后系统在 ESP 异常检测准确率上的整体提升幅度，并便于与其他基线模型对比，由于其他基线模型仅考虑工况数据输入。

为在相同输入条件下公平评估所提方法的有效性，本文在第一种数据集配置（仅工况输入）下，将所构建的多专家一仲裁式异常检测系统与多种基线模型进行对比，所选基线包括 XGBoost[19]、随机森林[20]（RF）、ResNet-50[21]以及 LSTM[22]。基线模型均采用公开库 sklearn 和 Pytorch 库中的标准模型结构。XGBoost 与随机森林将每个时间窗口内的二维工况样本展平为一维向量，直接进行二分类判别；LSTM 保留多变量时序结构，以二维时序矩阵作为输入；ResNet-50 则将输入整理为单通道二维张量，利用卷积网络提取局部时序模式及跨变量关联特征，最终输出故障诊断结果。

对于上述实验，考虑工业部署对推理成本的限制，本文将 ESP 异常检测专家模型集合的规模固定为 3。具体地，从 5 个基础模型中枚举所有 10 种可能的 3 模型组合，分别进行测试，以 F_1 最优的组合作为系统在两种输入条件下的最终配置。

实验同时考察了多专家模型集合的规模效应。在保持各训练设置一致的前提下，本文分别构建由 2～5 个中小参数模型组成的专家模型集合，并在统一的仲裁提示与输出结构约束下生成最终诊断结果，以考查系统异常检测性能随模型数量增加的变化趋势。在规模效应分析的基础上，本文进一步开展模型组合的对比实验。即在固定集合规模的前提下，评估不同中小参数 LLM 的组合方式对异常检测结果的影响，以检验模型来源异质性对系统性能的贡献程度。

2.3　结果分析

在仅工况输入条件下，本文所提出的系统与基线模型的异常检测结果如图 3 所示。

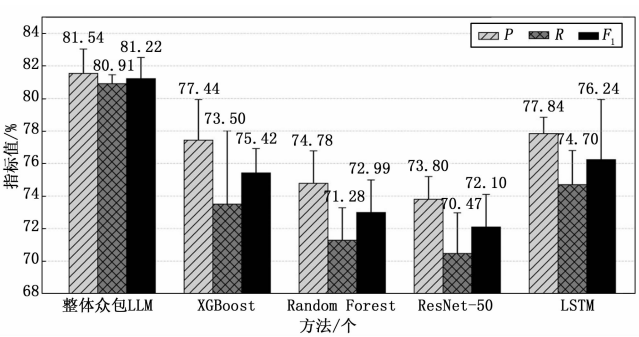

图 3　仅工况信息输入下不同方法的异常检测结果对比

从该组结果可以说明，本文所提出的 ESP 异常检测系统在不依赖额外的油井背景知识的条件下，P、R

与 F_1 分别达到 81.54%、80.91%与 81.22%，显著高于传统机器学习与深度模型基线。该结果表明，即便仅给定工况输入，LLM 仍能依托其较强的语义抽象与模式归纳能力，对异常相关的工况组合特征进行更有效的判别。从 3 项指标的相对关系看，多数基线方法呈现出 P 远高于 R 的趋势，说明在仅工况输入时，模型更倾向于保守预测以减少误报，但由此可能导致对异常样本的漏检。相比之下，本文所提出的基于大语言模型的异常检测系统能够同时维持较高 P 与 R，体现出对异常模式的更强鲁棒性与更稳健的决策边界。误差棒结果亦显示，其性能波动总体可控，而传统模型更易受数据分布变化影响。

在各设置一致的前提下，图 4 给出了两种输入设置下的异常检测的结果。

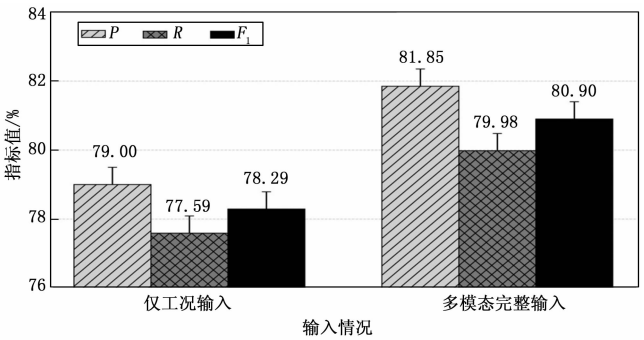

图 4 仅工况输入与完整输入下多专家—仲裁式异常检测系统性能对比

结果显示，包含了对应油井背景信息的完整上下文输入在 3 项指标上均取得一致提升：P、R 与 F_1 分别提升了 4.59%、2.53%和 3.07%。该增益表明，当模型在工况信息之外获得油藏地质、井史、设备配置等油井背景上下文时，能够更准确地分析工况数据，从而显著增强对异常模式的识别可靠性。

进一步观察可见，P 的提升幅度大于 R 的提升幅度，表明油井背景上下文对减少误报的作用更为突出。这可能是由于不同井型、不同设备组合与不同开发阶段可能对应不同的合理波动区间与操作策略，仅基于工况输入时，模型易将某些油井特定的工况偏移误判为异常。引入静态先验与历史信息后，LLM 能将当前工况观测置于更具物理与工程语义的上下文中进行归因与对比，进而更有效地区分“异常”与“井况差异或工艺差异导致的正常变化”，从而提升精确度并同步改善整体 F_1 的表现。

为评估候选模型数量对整体系统性能的影响，本文在相同训练数据、输入模板与 LoRA 配置下，分别构建由 2～5 个中小参数模型组成的候选集合，并保持仲裁提示与输出结构一致。图 5 汇总了不同专家模型集合规模下的异常检测结果。

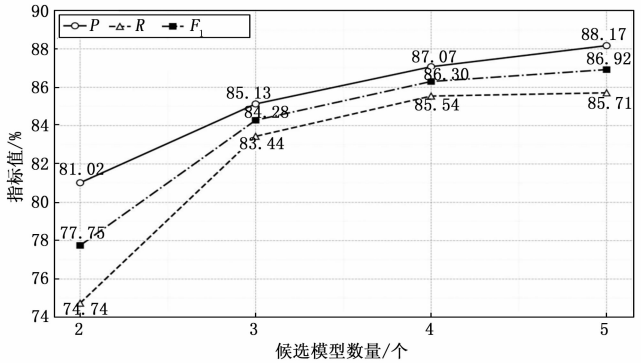

图 5 多专家候选集规模对 ESP 异常检测性能的影响

图中整体趋势表明，随着候选集合由 2 个扩展至 5 个模型，系统的 P、R 与 F_1 均呈单调上升的改善态势，说明多模型候选与仲裁机制能够有效缓解单一模型在工况理解与异常判别上的偶然偏差，从而提升整体决策的稳健性。具体而言，当候选数量从 2 增加到 3 时，3 项指标出现最显著的跃升；当候选数量进一步从 3 扩展至 4、5 时，性能仍有提升，但边际收益开始下降。这表明在候选规模达到一定程度后，新增模型带来的信息增量对分析多样性的拓展贡献有限。换言之，仲裁机制在 3 个模型时已能形成较稳定的互补结构，后续扩容不能提供明显的提升。在工程可用性角度，考虑推理成本与收益平衡，3～4 个候选模型可视为较优配置区间，既能获得接近上限的性能提升，又避免因候选过多导致的计算开销显著增长与仲裁复杂度上升。

在候选数量固定的前提下，使用不同中小参数 LLM 的组合的模型候选集的结果如表 2 所示。

表 2 不同专家模型组合的 ESP 异常检测性能对比 %

Qwen 2.5-7B	Mistral 7B	Llama-3.1-8B	GLM-4-9B-Chat	Intern LM2.5-7B	P	R	F_1
√	√	√	×	×	85.13	83.44	84.28
√	√	×	√	×	82.52	78.13	80.27
√	√	×	×	√	80.51	79.12	79.81
√	×	√	√	×	83.05	81.38	82.21
√	×	√	×	√	83.91	78.43	81.08
√	×	×	√	√	81.32	78.15	79.70
×	√	√	√	×	80.20	80.53	80.36
×	√	√	×	√	79.15	78.99	79.07
×	√	×	√	√	82.57	80.72	81.63
×	×	√	√	√	80.21	80.95	80.58
×	√	√	√	√	86.02	84.10	85.05
√	×	√	√	√	86.48	84.62	85.54
√	√	×	√	√	86.21	84.35	85.27
√	√	√	×	√	85.90	84.28	85.08
√	√	√	√	×	87.07	85.54	86.30
√	√	√	√	√	88.17	85.71	86.92

在候选数量为 3 的设置下，本文枚举了全部种组合。其中，Qwen2.5-7B＋Mistral 7B＋Llama-3.1-8B 的组合取得了该规模下最优结果，而包含 InternLM2.5-7B 的部分三模型组合整体略低。整体来看，不同组合的性能存在一定波动，F_1 的区间跨度约为 5.21%。当候选数量提升至 4 后，各组合性能显著收敛。其中，剔除 InternLM2.5-7B 的四模型组合（Qwen＋ Mistral＋Llama＋GLM）取得该组最优表现，并且区间波动降低，仅有 1.25%。进一步地，五模型全量组合达到最佳结果。

综上可得，在候选数量固定的前提下，不同中小参数 LLM 的组合确实会影响异常检测效果，随着候选集合变大，系统对单一模型优劣的敏感性降低，可能由于多模型带来的覆盖增益与互补效应[23-24]。

3　结束语

本文围绕潜油电泵（ESP）复杂工况下的异常检测任务，针对油井工况数据复杂，油井间分布漂移显著，以及异常漏检和错检等问题，提出了一种面向 ESP 场景的多专家－仲裁式协同大语言模型诊断方法。该方法将压力、电流与温度等多变量工况时序与油井地质、井史事件、设备配置等背景知识作为统一上下文进行联合推理，以 LoRA 方式对多种中小参数基座模型进行领域对齐，并引入 LLM-as-a-Judge 仲裁机制对多模型候选输出进行一致性校验与证据归纳，从而形成可核验的异常判别及解释性文本结果。实验结果表明，协同大模型系统在进一步引入油井背景知识，丰富上下文输入后，系统在各个指标上均获得提升。本文方法在 ESP 异常检测领域取得了积极的效果，但仍存在进一步提升空间，在时序信息与 LLM 的耦合方式上仍然存在提升空间，多模态上下文的结构化表达可进一步强化，未来可以探索更规范的知识组织方式，例如构建井级知识图谱，或可检索的结构化记忆库。面向在线部署场景，仍需在推理延迟与异常预警阈值策略等方面开展更细致的工程优化。总体而言，多模态、多专家－仲裁式协同大语言模型 ESP 异常诊断框架为油田在线监测与智能运维提供了一种兼具准确性与可解释性的系统方案。

参考文献：

[1] DUTTA N，KALIANNAN P，SUBRAMANIAM U. Application of machine learning algorithm for anomaly detection for industrial pumps [M]. Machine Learning Algorithms for Industrial Applications，Cham：Springer International Publishing，2020：237－263.

[2] 李　颜. 数据治理在大庆油田智能运维中的应用研究 [J]. 中国管理信息化，2025，28（14）：163－165.

[3] 周子文，忻　达. 基于 DBN 的油田 SCM 液压系统故障诊断研究 [J]. 液压气动与密封，2025，45（7）：107－112.

[4] 薛　冰. 基于深度学习的抽油机工况诊断模型研究 [D]. 大庆：东北石油大学，2025.

[5] PENG L，HAN G，PAGOU A L. Real-time anomaly detection methodology in the electric submersible pump systems [J]. Petroleum Science and Technology，2022，40（21）：2669－2684.

[6] BHARDWAJ A S，SARAF R，NAIR G G，et al. Real-time monitoring and predictive failure identification for electrical submersible pumps [C] //Abu Dhabi International Petroleum Exhibition and Conference，SPE，2019：D021S049R003.

[7] LIN B，JIN Y，CAO Q，et al. Developing a large language model for oil-and gas-related rock mechanics：Progress and challenges [J]. Natural Gas Industry B，2025，12（2）：110－122.

[8] 杨寒雨，赵晓永，王　磊. 数据归一化方法综述 [J]. 计算机工程与应用，2023，59（3）：13－22.

[9] 李荣涵，浦荣成，沈佳楠，等. 基于思维链的大语言模型知识蒸馏 [J]. 数据采集与处理，2024，39（3）：547－558.

[10] WANG Y，WU S，ZHANG Y，et al. Multimodal chain-of-thought reasoning：a comprehensive survey [J]. Arxiv Preprint Arxiv：2503.12605，2025.

[11] LIU A，FENG B，XUE B，et al. Deepseek-v3 technical report [J]. Arxiv Preprint Arxiv：2412.19437，2024.

[12] 巴泽智，张　辉，谢铮涵，等. 大语言模型自动化提示工程技术研究综述 [J]. 计算机科学与探索，2025，19（12）：3131－3152.

[13] YANG A，YANG B，HUI B，et al. Qwen2.5 technical report [J]. Arxiv Preprint Arxiv：2412.15115，2024.

[14] JIANG A Q，SABLAYROLLES A，MENSCH A，et al. Mistral 7B [J]. Arxiv Preprint Arxiv：2310.06825，2023.

[15] GRATTAFIORI A，DUBEY A，JAUHRI A，et al. The Llama 3 herd of models [J]. Arxiv Preprint Arxiv：2407.21783，2024.

[16] GLM Team，ZENG A，LIU B，et al. ChatGLM：a family of large language models from GLM-4 All Tools [J]. Arxiv Preprint Arxiv：2406.12793，2024.

[17] CAI Z，CAO M，CHEN H，et al. InternLM2 Technical Report [EB/OL]. （2024－03－26）[2025－05－19]. https://arxiv.org/pdf/2403.17297.pdf.

[18] HU E J，SHEN Y，WALLIS P，et al. Lora：low-rank adaptation of large language models [J]. ICLR，2022，1（2）：3.

（下转第 24 页）