

电力物资供应链电子数据信任强度自适应分级方法

俞春华¹, 郭庆雷^{2,3}, 杨珂⁴, 柏能¹, 黄宜华⁵

1. 国网江苏省电力有限公司物资分公司, 江苏 南京 210036;
2. 国网区块链科技(北京)有限公司, 北京 100077;
3. 国网区块链应用技术实验室, 北京 100077;
4. 国网数字科技控股有限公司, 北京 100077;
5. 南京大学计算机软件新技术国家重点实验室, 江苏 南京 210023

摘要

电力物资供应链电子数据管理需求复杂, 传统信任强度分级方法受限于静态规则与单一评价维度, 缺乏领域知识引导, 难以满足精细化与差异化的应用要求。针对这一问题, 提出一种电力物资供应链电子数据信任强度自适应分级方法——Adap-LLM。该方法将专家构建的多维评分体系与分级规则以结构化提示词的形式注入大语言模型, 并结合参数高效微调实现领域知识深度融合, 从而实现对多种类型电力文档信任强度的自适应分级。实验结果表明, Adap-LLM 在私有电力供应链数据集与基于公开数据源构建的跨领域评估数据集上的各项指标均优于对比模型, 充分验证了其在电力物资大数据分级管理中的有效性与先进性, 具有良好的工程应用价值。

关键词

供应链电子数据分级; 大语言模型; 领域知识; 提示词工程; 参数高效微调

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.2096-0271.BDR26030

Adaptive grading method of trust intensity for electronic data in the electric power material supply chain

Yu Chunhua¹, Guo Qinglei^{2,3}, Yang Ke⁴, Bai Neng¹, Huang Yihua⁵

1. State Grid Jiangsu Electric Power Co., Ltd., Materials Branch, Nanjing 210036, China
2. State Grid Blockchain Technology (Beijing) Co., Ltd., Beijing 100077, China
3. State Grid Blockchain Application Technology Laboratory, Beijing 100077, China
4. State Grid Digital Technology Holding Co., Ltd., Beijing 100077, China
5. State Key Laboratory for Novel Software Technology at Nanjing University, School of Computer Science, Nanjing 210023, China

Abstract

The management of electronic data in power material supply chains is characterized by complex requirements, while traditional trust strength classification methods are constrained by static rules and single evaluation dimensions, and lack sufficient domain knowledge guidance, making it difficult to meet the demands of fine-grained and differentiated

applications. To address this issue, an adaptive trust strength classification method for electronic data in power material supply chains, named Adap-LLM, was proposed. In this method, a multi-dimensional scoring system and classification rules constructed by domain experts were injected into a large language model (LLM) in the form of structured prompts, and parameter-efficient fine-tuning was integrated to achieve deep fusion of domain knowledge, thereby enabling adaptive classification of trust strength for various types of power documents. Experimental results demonstrated that Adap-LLM outperforms the compared models across all evaluation metrics on both private power supply chain dataset (PPSCD) and a cross-domain evaluation dataset constructed from public data sources (CDE-PDS), which fully validates its effectiveness and advancement in the hierarchical management of big data in power material supply chains, and exhibits promising engineering application value.

Key words
supply chain electronic data grading, LLM, domain knowledge, prompt engineering, parameter-efficient fine-tuning

0 引言

随着能源互联网与电力供应链数字化的深度融合，电力物资供应链中的电子数据总量与类型持续增长，结构日趋复杂。对其可信性进行分级管理，直接关系到电网物资管理的审计合规性、合法性及运营效率。不同类别的电力物资供应链电子数据在敏感性、法律效力和业务影响程度上存在差异。这种差异性要求在数据存储、查询与溯源的过程中，根据数据的信任强度等级实施自动化分级管理与差异化信息保护。

在有关电子数据自动分级与可信评估的研究中，现有方法主要分为规则驱动方法、深度语义建模方法和可信数据治理方法3类。规则驱动方法通常依赖人工设定规则、统计特征、概率推理或源可靠性估计实现数据分类与可信度判断，具备一定的可解释性，但对规则设计和人工特征选取依赖较强，难以适应复杂业务文档中的语义变化^[1-2]。深度学习方法通过多标签神经网络、分层图卷积和长文档建模机制，增强模型对文本语义的表示能力，能在一

定程度上缓解传统方法对人工特征的依赖，但对法律风险、业务价值、审计追溯等领域规则的建模仍不够充分^[3-5]。面向多源异构数据和工业场景的研究进一步引入跨模态融合、领域本体和动态信任管理机制，为复杂数据分类、可追溯治理和多维可信评估提供了有益参考^[6-10]。总体而言，现有方法在电子数据分级中仍存在评价维度单一、规则体系静态、领域知识与语义模型融合不足等问题，难以满足电力物资供应链电子数据对精细化、差异化和自适应分级的需求。

针对上述问题，本文提出一种基于大语言模型的电力物资供应链电子数据信任强度自适应分级方法——Adap-LLM。该方法以四维评分体系和C/P/R分级规则为知识基础，将结构化提示词嵌入评分维度、分级阈值、推理步骤和输出约束，并结合低秩适配（low-rank adaptation, LoRA）微调，学习文档语义特征与信任强度等级之间的映射关系，从而实现电子数据的自适应分级。

本文的主要贡献如下：
(1) 构建了面向电力物资供应链电子数据的四维评分体系和C/P/R分级标准，为电子数据差异化管理提供规则依据；

(2) 提出融合结构化提示词与 LoRA 微调的 Adap-LLM 方法, 实现领域规则与大语言模型语义理解能力的协同建模;

(3) 在私有电力物资供应链数据集 (private power supply chain dataset, PPSCD) 和基于公开数据源的跨领域评估数据集 (cross-domain evaluation dataset based on public data sources, CDE-PDS) 上开展对比实验、消融实验和权重敏感性分析, 验证了 Adap-LLM 的有效性和稳健性。

1 任务描述

在电力物资供应链电子数据信任强度自适应分级任务中, 电力供应链中的电子文档集合为 $\mathbf{D}=\{d_1, d_2, \dots, d_n\}$, 单个文档 d_i 可表示为:

$$d_i=\{T_i, M_i, C_i\} \quad (1)$$

其中, T_i 表示文档主要文本内容, M_i 表示文档类型、文档名、文档路径等元数据信息, C_i 表示文档所处业务环节、流程节点及上下文语义信息。

任务目标是学习映射函数 f :

$$f:\mathbf{D} \times \mathbf{K} \rightarrow L_1 \times L_2 \quad (2)$$

其中, $\mathbf{K}$ 表示领域专家知识集合, 包括评分标准、分级阈值、典型案例等专家经验; $L_1=\{C, P, R\}$ 表示一级分类标签; $L_2=\{C_i, P_j, R_k | i, j, k \in \{1, 2, 3\}\}$ 表示二级分类标签。

本文以 C/P/R 一级信任强度等级为主要分级目标, 二级分类用于刻画各一级等级内部的业务属性差异, 并为后续领域知识注入、模型推理约束及解释生成提供细化依据。

为保证分级规则具有明确的业务依据,

本文依据国家电网数据管理规范、物资供应链业务制度、合同履行管理要求及审计合规要求, 构建面向电力物资供应链电子数据的多维度量化评分框架。该框架由国网物资领域专家组结合合同签订、招标采购、履约验收等典型业务环节归纳形成, 重点刻画不同电子文档在法律效力、业务控制、审计追溯和敏感保护等方面的作用差异。在此基础上, 从法律效力、业务控制、审计追溯和敏感保护 4 个方面提炼出法律风险程度、业务重要性、审计合规要求和数据敏感性 4 个核心维度。

设 $s_j(d_i)$ 表示文档 d_i 在第 j 个评分维度上的得分, w_j 表示评分维度的对应权重, 则文档 d_i 的综合信任强度评分 $S(d_i)$ 可表示为:

$$S(d_i)=\sum_{j=1}^4 w_j \cdot s_j(d_i) \quad (3)$$

其中, 法律风险程度、业务重要性、审计合规要求和数据敏感性的权重分别设为 35%、30%、20% 和 15%。该权重设置综合考虑了法律效力、业务控制、审计追溯和敏感保护在电力物资供应链数据治理中的相对重要性, 并经领域专家讨论确定。

基于上述综合评分结果, 本文构建 C/P/R 三级信任强度分级标准, 见表 1。在一级分类基础上, 进一步设置二级分类细则, 用于描述同一信任强度等级内部的业务属性差异, 具体见表 2。

2 基于 LLM 的电力物资供应链电子数据信任强度自适应分级方法

针对上述问题, 本文提出基于 LLM 的电力物资供应链电子数据信任强度自适应分级方法—Adap-LLM。该方法以四维评

分体系、C/P/R 分级标准及二级分类细则为领域知识基础，通过结构化提示词将专家规则、评分依据、推理步骤和输出格式嵌入模型输入，并结合 LoRA 参数高效微调，学习电力物资供应链场景下文档语义特征与信任强度等级之间的映射关系。Adap-LLM 方法整体架构如图 1 所示。

2.1 领域知识注入的递进式架构设计

领域知识注入的目标是将电力物资供应链电子数据分级所涉及的业务规则、评分标准和专家经验，转化为可被 LLM 理解和遵循的推理约束，从而提升 LLM 在专业场景下的判别一致性与结果可解释性。针对电力物资供应链电子数据在法律效力、业务控制、审计追溯和敏感保护等方面的差异，本文构建递进式领域知识注入架构，

如图 2 所示。

该架构按照“角色限定—规则导入—任务分解—案例对齐—推理约束—格式规范”的逻辑组织领域知识，具体包括系统层、专家知识层、任务层、案例层、推理层和输出层。系统层用于限定模型的专业角色和任务场景；专家知识层引入四维评分体系、C/P/R 分级标准及二级分类规则；任务层将分级过程分解为文档特征识别、多维度评分、综合计算和等级判定；案例层通过典型样例建立文档特征、评分依据与分级结果之间的对应关系；推理层围绕文档类型、业务环节、法律效力、审计价值和敏感程度等要素约束模型的推理分析路径；输出层统一模型生成结果的字段结构和表达形式。

通过上述分层设计，静态分级规则被组织为贯穿输入、推理和输出过程的知识

表1 电子数据信任强度一级分类标准与阈值

一级分类	评分阈值	分类特征	说明
C 级	$S(d_i) \geq 75$	决定供应链成败的核心文档,具有不可替代性和决定性影响	对供应链关键决策起到决定性作用
P 级	$50 \leq S(d_i) < 75$	维护供应链流程运转的关键文档,具有重要的过程控制价值	在业务流程中具有重要的控制功能
R 级	$S(d_i) < 50$	记录和支撑业务操作的辅助文档,主要起备查和追溯作用	辅助支撑和备查追溯

表2 电子数据信任强度二级分类细则

一级分类	二级分类	分类条件	典型文档示例
C 级	C1 级(法律强制类)	法律风险程度 ≥ 85 分	正式合同、中标通知书
	C2 级(财务决算类)	业务重要性 ≥ 85 分	重大付款凭证、结算文件
	C3 级(履约确认类)	审计合规要求 ≥ 85 分	验收报告、履约证明
P 级	P1 级(技术规范类)	技术复杂度高、专业性强	技术规范书、专家意见
	P2 级(商务流程类)	商务流程关键节点	招标文件、评标报告
	P3 级(管理控制类)	管理控制重要环节	质量报告、风险评估
R 级	R1 级(数据统计类)	格式标准化	货物清单、统计报表
	R2 级(流程记录类)	时间序列特征	会议记录、操作日志
	R3 级(申请通知类)	交互传达功能	申请表单、通知公告

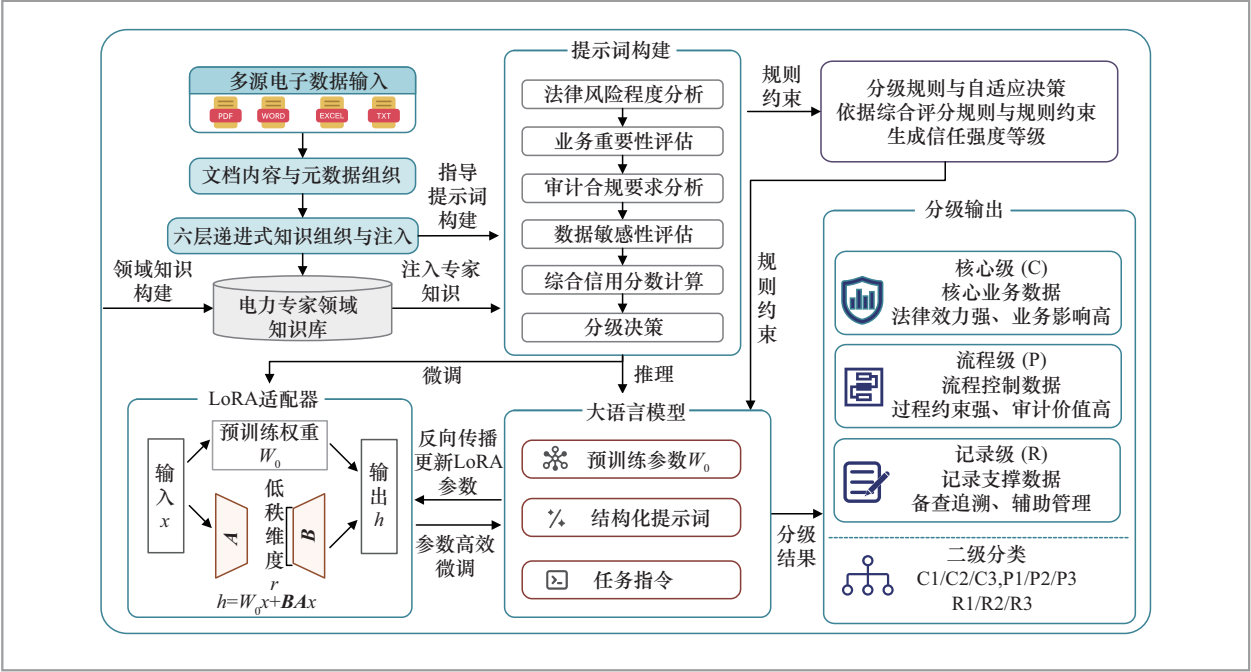


图1 Adap-LLM方法整体架构

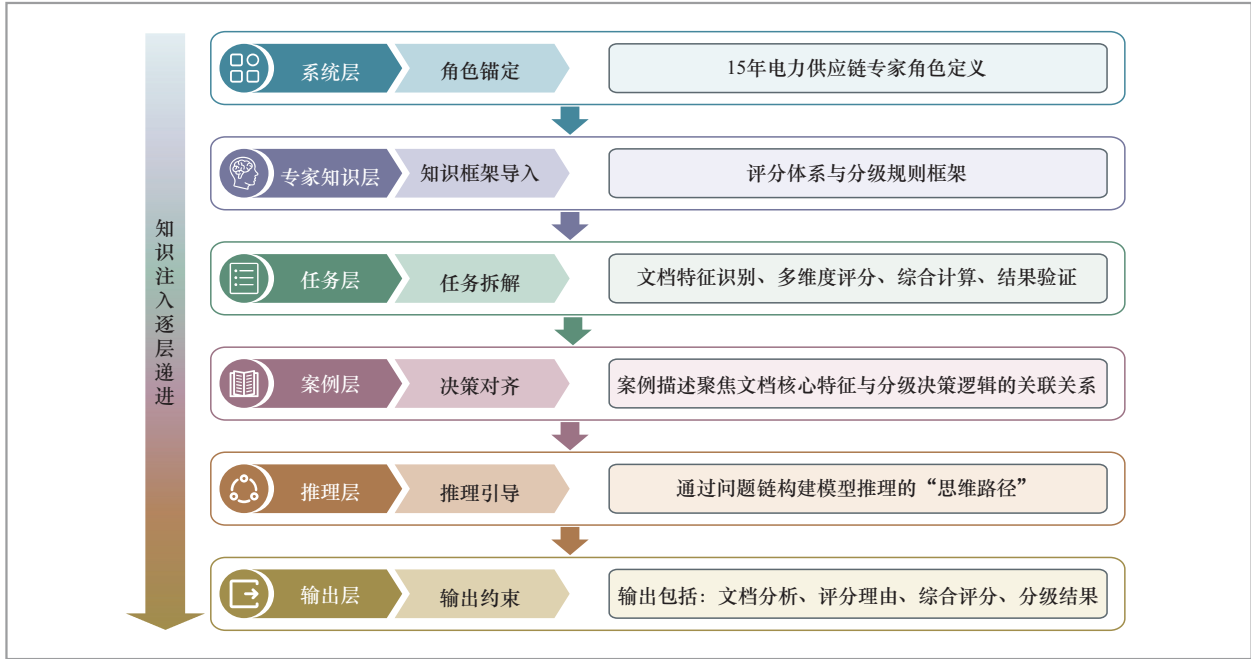


图2 递进式领域知识注入架构

约束链，使LLM能够在发挥通用语义理解能力基础上融合电力物资供应链领域规则，为后续结构化提示词构建和LoRA参数高效微调提供一致的知识基础。

2.2 结构化提示词模板构建

在递进式领域知识注入架构的基础上，进一步构建结构化提示词模板，将四维评分体系、C/P/R 分级标准、二级分类细则、推理步骤和输出格式统一封装为模型输入。该模板由系统层提示词和用户层提示词两部分构成，其构建框架如图 3 和图 4 所示。

系统层提示词主要承担领域规则注入功能，包括角色定义、核心任务、评分体系和输出约束。其中，角色定义用于设定模型的专业身份；核心任务明确模型需完成的电子数据信任强度分级目标；评分体系嵌入法律风险程度、业务重要性、审计合规要求和数据敏感性 4 个维度及其对应权重；输出约束规定模型需按照统一结构生成文档特征、评分依据、综合得分和分级结果。

用户层提示词主要承担任务实例化功能，包括推理引导、待分析内容和案例提示模块。推理引导用于约束模型按照“文档类型识别—关键特征提取—四维评分—综合得分计算—等级判定”的顺序完成分析，该设计通过显式化中间分析步骤增强模型对复杂分级任务的推理能力^[11]；待分析内容提供文档文本、元数据及业务上下文信息；案例指示通过典型样例增强模型对分级边界和业务语义的理解。

结构化提示词模板参数示例见表 3。其中，在模板的 System 部分，Instruction 字段明确声明了角色信息、专家知识、任务约束等内容；Input 部分提供了具体的分析步骤要求、待评估文档，以及案例指示。同时，该模板还规定了 LLM 的类型及其相应输入输出的界定符格式。

经过结构化提示词模板封装后，面向指令微调的模型输入组织结构如图 5 所示。模型输入由任务指令（Instruction）

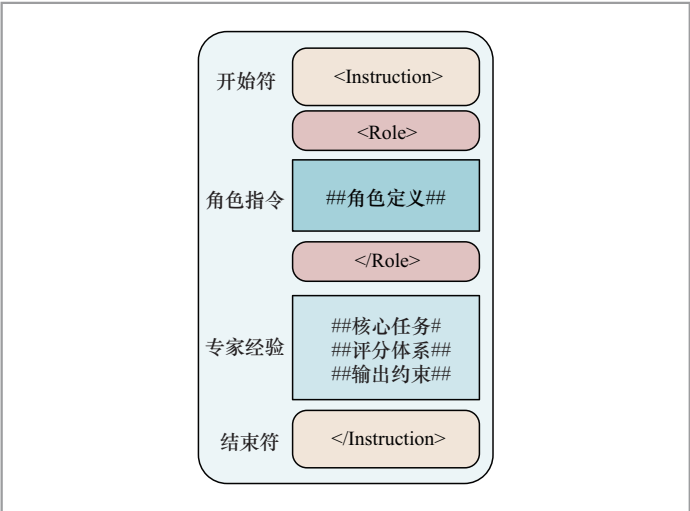


图 3 系统层提示词构建框架

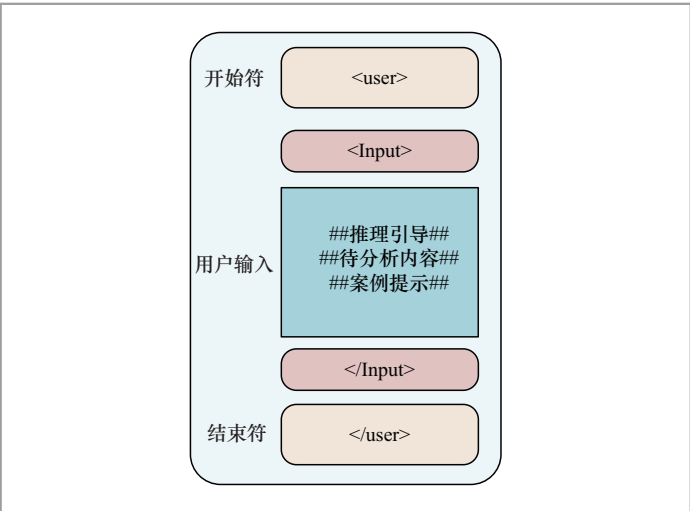


图 4 用户层提示词构建框架

和待处理模板（Text）两部分组成，前者用于承载电力物资供应链电子数据的业务属性、评分维度、分级规则和输出约束，后者用于提供待分析文档的内容及其元数据信息。模型输出则包含文档特征、评分依据、综合得分和分级结果的结构化回复。通过上述组织方式，提示词不仅向模型提供分类任务指令，还将专家规则、推理路径和输出格式统一纳入输入上下

表3 结构化提示词模板参数

参数名	参数值
System	< im_start >instruction\n
System	< im_end >\n
user_start	< im_start >Input\n
user_end	< im_end >\n
LLM_start	< im_start >assistant\n
LLM_end	< im_end >\n

文，使模型能够在领域规则约束下完成信任强度分级。

2.3 面向分级任务的模型微调方法

为进一步提升模型对电力物资供应链场景的适配能力，在结构化提示词基础上引入 LoRA^[12]参数高效微调方法。该方法可冻结基座大语言模型主体参数，仅在注意力层和前馈网络中引入 LoRA 矩阵，通过少量可训练参数学习电力物资文档特征、领域规则与信任强度标签之间的映射关系，从而在降低训练成本的同时保持模型原有的语义理解能力。

设基座模型中的待适配权重为 W_0 ，且其在训练过程中保持冻结，LoRA 通过低秩矩阵 A 、 B 对权重增量进行近似，则该层前向传播的输出 h 可表示为：

$$h = W_0x + BAx \tag{4}$$

其中， x 表示模型输入。

该机制使模型能够在不进行全参数更新的前提下，学习电力物资供应链电子数据分级任务中的领域差异性特征。

采用指令监督微调范式构建训练样本，输入序列由结构化提示词、待分析文档及领域分级规则组成，目标输出包括文档关键特征、四维评分、评分依据、综合得分、一级分类和二级分类等字段。训练目标是在给定输入条件下最大化标准输出序列的生成概率，其损失函数可表示为：

$$L = - \sum_{t=1}^T \log P_{\theta}(y_t | y_{<t}, X) \tag{5}$$

其中， θ 表示 LoRA 引入的可训练适配参数， X 表示模型输入序列， T 表示目标输出序列的长度， y_t 表示目标输出序列中第 t 个 Token， $y_{<t}$ 表示第 t 个 Token 之前已生成的 Token 序列， P 表示条件概率。

经过该训练过程，模型不仅能学习文档内容与 C/P/R 等级之间的对应关系，还能够进一步内化提示词中嵌入的评分规则、推理步骤和输出格式约束。最终，本文所提方法可形式化表示为：

$$\text{Label} = \text{LLM}_{\text{LoRA}}(P_{\text{Template}}, \text{Doc}) \tag{6}$$

其中， P_{Template} 表示结构化领域知识提示词，Doc 表示待分级电子文档， LLM_{LoRA} 表示经 LoRA 微调后的 LLM，Label 表示模型生成的信任强度分级结果。

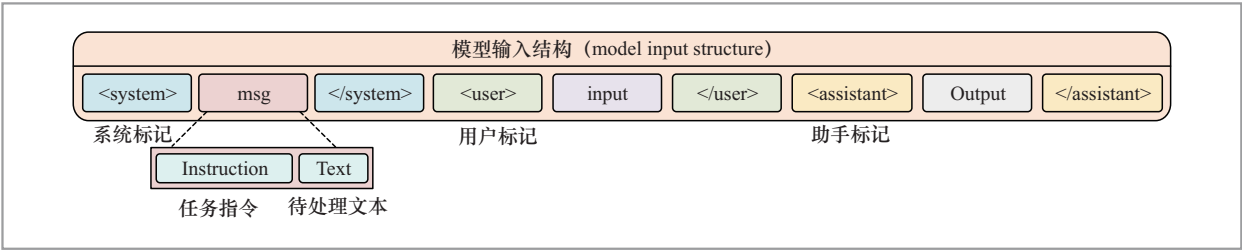


图5 面向指令微调的模型输入组织结构

3 实验

3.1 数据集

本文实验基于 PPSCD 和 CDE-PDS 展开验证。PPSCD 源于电力物资供应链真实业务场景，覆盖合同签订、招投标、履约验收等关键业务环节。原始数据共包含 3 856 份异构电子文档，格式包括 PDF、Word、Excel 和 TXT 等，典型文档类型包括正式合同、技术规范书、招标文件、验收报告、会议记录、统计报表等。为保证数据质量，本文对原始文档进行统一文本抽取、噪声清洗、隐私脱敏和关键内容提取，形成用于模型训练和评估的核心文本内容。

在数据标注阶段，本文采用“双盲标注—一致性核查—争议复核”流程生成数据集标签。具体而言，由未参与评分体系构建的国网物资领域专家依据四维评分体系和 C/P/R 分级规则对全部样本进行双人独立标注；对于标签不一致、四维评分差异超过设定阈值或综合得分接近 C/P/R 分界点的样本，引入额外专家复核裁定机制，并以复核结果为标准标签。标注规则的构

建、样本标注与训练提示词设计相互分离，以降低评分规则设计与样本标注之间的同源偏置风险。考虑到电力物资供应链电子文档普遍篇幅较长，单份文档中通常包含多个能够反映其信任强度的业务片段，本文在保持专家标签不变的前提下，对原始文档进行片段级样本构建，将标题、关键条款及金额等内容组织为模型输入，并与对应专家标签组合，形成“指令—输入—输出”的结构化微调样本。最终，3 856 份原始电子文档被构建为 18 650 条结构化样本。

为验证 Adap-LLM 方法的跨领域泛化能力，本文进一步基于公开数据源构建跨领域评估数据集 CDE-PDS。该数据集融合 ChnFinAnn、CAIL 和 THUCNews 3 个公开中文数据源，分别覆盖金融公告、司法文书和新闻文本。本文依据 C/P/R 分级规则对数据进行标签映射：将具有强法律效力或高风险约束属性的司法文书、股权冻结公告及股权质押公告映射为 C 级；将反映企业关键业务流程和监管披露要求的股权回购、股权增持公告及股权减持公告映射为 P 级；将主要承担信息传递、事实记录和备查参考功能的财经简讯、事务性通知及一般新闻文本映射为 R 级。

PPSCD 与 CDE-PDS 的数据分布见表 4。为分析模型在临近分级阈值样本上

表4 PPSCD与CDE-PDS的数据分布

数据集	分类	文档类型	数量/条
PPSCD	C	合同、中标通知书、结算文件、验收报告等	6 133
PPSCD	P	技术规范书、招标文件、评标报告、质量报告等	7 322
PPSCD	R	会议记录、统计报表、货物清单、通知公告等	5 195
合计	—	—	18 650
CDE-PDS	C	司法文书、股权冻结公告、股权质押公告	4 000
CDE-PDS	P	股权回购公告、股权增持公告、股权减持公告	4 000
CDE-PDS	R	财经简讯、事务性通知、一般新闻文本	4 000
合计	—	—	12 000

的判别能力，本文进一步对PPSCD中的边界样本进行统计。根据表1所示C/P/R分级阈值，将综合评分位于[58,62]的样本定义为R/P边界样本，将综合评分位于[73,77]的样本定义为P/C边界样本。PPSCD边界样本分布统计见表5。

3.2 评测标准

本文使用准确率（accuracy，Acc）、精确率（precision，Pre）、召回率（recall）和F1分数作为电力物资供应链电子数据信任强度自适应分级任务的评估指标。准确率表示预测类别与真实类别一致的样本数占总样本数的比例。精确率、召回率和F1分数采用One-vs-Rest策略，分别对各类别进行计算，即在计算某一类别指标时，将该类别视为正类，其余类别视为负类。其中，精确率表示被预测为该类别的样本中实际属于该类别的比例，召回率表示实际属于该类别的样本被正确预测为该类别的比例，F1分数为精确率与召回率的调和平均值。本文进一步对各类别的精确率、召回率和F1分数进行宏平均，作为最终的精确率、召回率和F1分数指标。各类别指标计算所涉及的TP、FP、TN和FN定义见表6。

基于上述定义，各指标的数学表达式如下：

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Pre = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{10}$$

3.3 实验设置

本文将Qwen3-4B作为主要基座模型，并引入LLaMA、DeepSeek和ChatGLM3开展对比实验。实验环境为Linux操作系统，配备两台NVIDIA T4-16G GPU服务器，编程语言为Python 3.10.8，深度学习框架为PyTorch 2.6。

为避免同一原始文档的不同片段同时出现在训练集和测试集中，本文先在原始文档层面按照8：1：1划分训练集、验证集和测试集，再分别对各子集内的文档进行片段级样本构建，并引入分层抽样策略，以确保各类别样本比例基本一致。模型微调采用LoRA方法，冻结基座模型全部参数，仅更新LoRA适配参数。LoRA作用模块包括q_proj、k_proj、v_proj、o_proj、gate_proj、up_proj和down_proj；秩参数r设为16，缩放系数α设为32，Dropout设为0.05。优化器采用AdamW，

表5 PPSCD边界样本分布统计

数据集	样本类型	C类/条	P类/条	R类/条	合计/条	占PPSCD比例
PPSCD	R/P边界样本	0	810	740	1 550	8.31%
PPSCD	P/C边界样本	780	910	0	1 690	9.06%
PPSCD	全部边界样本	780	1 720	740	3 240	17.37%
PPSCD	非边界样本	5 353	5 602	4 455	15 410	82.63%
合计	—	6 133	7 322	5 195	18 650	100.00%

学习率设置为 5×10^{-5} ，最大文本长度设置为2 048。受显存限制，模型训练批次大小为1，推理温度系数设为0.1，以降低生成随机性并提高输出稳定性。

本文以C/P/R一级信任强度等级为主要评价对象，其中C级、P级和R级分别对应核心级、流程级和记录级电子数据，用于表征电力物资供应链电子数据的主体信任强度等级；二级分类作为各一级等级下的细化子类，用于刻画同一等级内部的业务属性差异，并为领域知识注入、推理约束及结果解释生成提供细粒度规则依据。实验过程中，通过规则匹配抽取模型输出中的C/P/R一级分类字段，并与人工标注的主体信任强度等级进行比对，进而计算各项评价指标。实验所用训练提示词示例如图6所示。该示例将专家定义的四维评分体系、分级阈值，以及推理逻辑显式地注入模型的输入上下文。

3.4 对比实验

为验证 Adap-LLM 方法的有效性与跨模型适配能力，本文在 CDE-PDS 和 PPSCD 两个数据集上选取传统模型、预训练模型、纯大模型方法与 Adap-LLM 方法进行对比实验。其中，传统的预训练基线包括 Longformer^[13]、LSTM^[14]、BiLSTM-Attention^[15] 和 MacBERT^[16]，分别用于表征序列建模、注意力增强序列建模、中文预训练语义建模及长文本建模能力。纯大模型方法采用 Prompt-only 设置，在不进行参数微调的情况下，调用 LLaMA^[17]、DeepSeek-LLM^[18]、ChatGLM3^[19] 和 Qwen3^[20] 系列中可获得的高性能应用程序编程接口（application programming interface，API）版本，并输入完整结构化提示词完成推理。Adap-LLM 方法则选

表6 分类任务评估指标定义

	预测为正类	预测为负类
实际正类	TP	FN
实际负类	FP	TN

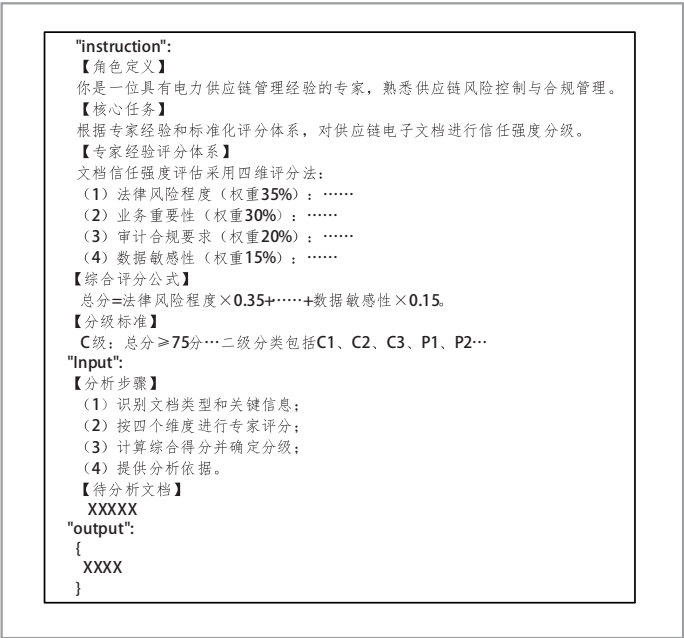


图6 训练提示词示例

取与上述模型系列对应的可本地部署轻量基座模型，包括 LLaMA-7B、DeepSeek-LLM-7B、ChatGLM3-6B、Qwen3-4B，在其基础上引入领域知识注入与 LoRA 微调，构建 Adap-LLM 系列模型。各模型在 CDE-PDS 与 PPSCD 数据集上的实验结果见表7。

由表7可知，传统模型整体性能相对有限，引入双向建模与注意力机制后，BiLSTM-Attention 相较于 LSTM 取得了一定的性能提升，说明注意力机制有助于捕捉文档中的关键片段。相比之下，MacBERT 和 Longformer 进一步提升了分级性能，表明预训练语义表示和长文本建模能力对电力物资供应链电子文档分级具

表7 各模型在CDE-PDS与PPSCD数据集上的实验结果

模型	CDE-PDS				PPSCD			
	准确率	精确率	召回率	F1分数	准确率	精确率	召回率	F1分数
传统模型								
LSTM	79.54%	78.12%	77.65%	77.88%	86.74%	85.92%	84.53%	86.15%
BiLSTM-Attention	83.41%	82.76%	81.94%	82.35%	89.06%	88.42%	87.95%	88.18%
MacBERT	86.34%	85.71%	85.16%	85.43%	90.72%	90.18%	89.64%	89.91%
Longformer	87.21%	86.45%	85.92%	86.18%	91.25%	90.47%	89.81%	90.83%
大模型方法(Prompt-only)								
LLaMA	87.86%	87.24%	86.71%	86.91%	91.82%	91.26%	90.74%	91.00%
DeepSeek-LLM	89.42%	88.95%	88.36%	88.65%	93.18%	92.64%	92.11%	92.37%
ChatGLM3	88.73%	88.14%	87.62%	87.88%	92.56%	92.07%	91.48%	91.77%
Qwen3	90.16%	89.68%	89.13%	89.40%	93.74%	93.21%	92.76%	92.98%
本文所提方法(Adap-LLM+不同大模型基座)								
LLaMA-7B	91.18%	90.62%	90.21%	90.41%	94.63%	94.08%	93.66%	93.87%
DeepSeek-LLM-7B	92.03%	91.46%	91.18%	91.32%	95.46%	94.92%	94.37%	94.64%
ChatGLM3-6B	91.56%	91.02%	90.85%	90.93%	95.12%	94.58%	94.25%	94.41%
Qwen3-4B	92.45%	91.88%	91.67%	91.77%	95.83%	95.20%	94.61%	95.48%

有积极作用。大模型方法（Prompt-only）整体优于传统模型，说明 LLM 在强提示条件下已具备较好的语义理解和规则推理能力。然而，由于 Prompt-only 未经领域任务微调，其对电力物资供应链业务语境和边界样本的适配能力仍然有限。本文所提方法在不同基座模型上均较对应 Prompt-only 基线取得提升。其中，以 Qwen3-4B 为基座的 Adap-LLM_{Qwen3-4B} 取得最佳结果，在 PPSCD 数据集上的 F1 分数达到 95.48%，较 Qwen3 Prompt-only 基线提升 2.50 个百分点；在 CDE-PDS 数据集上的 F1 分数达到 91.77%，提升 2.37 个百分点。该结果表明，在规则约束下的中文业务文档分级任务中，模型性能不仅取决于参数规模，还与中文语义理解能力、指令遵循能力、结构化输出稳定性，以及领域提示模板适配程度密切相关。上述结果验

证了结构化领域知识注入与 LoRA 微调能够进一步增强模型对分级规则、业务语义和输出格式的适应能力，且该方法不依赖特定基座模型，具有较好的跨模型迁移性。此外，各模型在 PPSCD 数据集上的表现普遍高于 CDE-PDS，主要由于 PPSCD 数据集与本文任务场景和标签规则更加一致，而 CDE-PDS 数据集包含金融公告、司法文书和新闻文本等跨领域数据，语义分布更复杂，分级难度更高。

大模型训练与微调的效率不仅取决于模型参数量，还受算力系统、并行加速策略、存储系统，以及容错机制等基础设施的影响。大模型基础设施建设应兼顾性能、可靠性和可扩展性，其中训练过程中的检查点访问性能直接关系到系统恢复效率与资源利用率^[21-22]。因此，在模型分级性能之外，本文进一步统计大模型方法

(Prompt-only) 和本文所提方法 (Adap-LLM+不同大模型基座) 的 token 消耗、端到端平均推理时间和输出吞吐量指标, 以及本文所提方法 (Adap-LLM+不同大模型基座) 的峰值显存占用, 如图 7~图 10 所示, 以从系统开销角度评估本文所提方法的部署性。其中, API 模型的峰值显存不可观测; 本地 Adap-LLM 模型部署于 NVIDIA T4 环境, 显存占用由推理过程记录获得。

由图 7~图 10 可知, Adap-LLM 在不同大模型基座下的平均输入、输出 token 数量整体低于 Prompt-only 方法, 表明其能够以较低的 token 开销完成分级任务。在推理效率方面, Adap-LLM 的端到端平均推理时间整体较低, 其中 Qwen3-4B 基座下的推理时间最低。在输出吞吐量方面, Prompt-only 方法采用 API 调用, Adap-LLM 部署于本地 NVIDIA T4 环境, 两者在计算资源和推理服务配置方面存在差异, 反映不同部署条件下系统运行存在开销差异。峰值显存结果表明, 不同基座模型的资源占用存在差异, 其中 Qwen3-4B 基座的显存占用相对较低。

3.5 消融实验

为验证领域知识注入、分步推理引导与 LoRA 微调对 Adap-LLM 方法分级性能的影响, 本文以 Qwen3-4B 为基座模型, 在 CDE-PDS 与 PPSCD 两个数据集上开展消融实验。各消融实验模型设置如下。

(1) Base: 未微调 Qwen3-4B, 仅输入基本任务说明, 不引入领域知识和结构化约束;

(2) Knowledge Prompt: 在 Base 的基础上仅加入四维评分体系、C/P/R 分级规则和结构化输出约束, 但不引入分步推

理引导、典型案例示范和 LoRA 微调, 用于验证基础领域规则提示对模型分级性能的影响。

(3) LoRA w/o Knowledge: 采用不含四维评分体系和 C/P/R 显式规则的训练样本进行 LoRA 微调, 但保留统一标签输出格式, 仅学习文档内容与分级标签之间的映射关系。

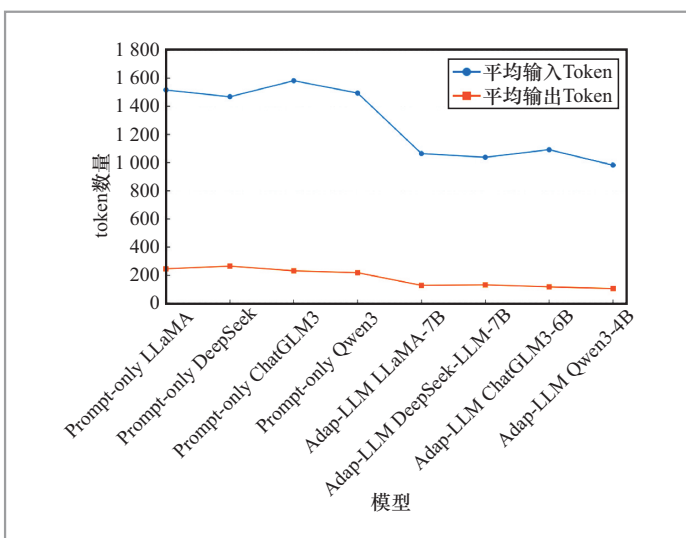


图7 不同方法下模型的Token消耗对比

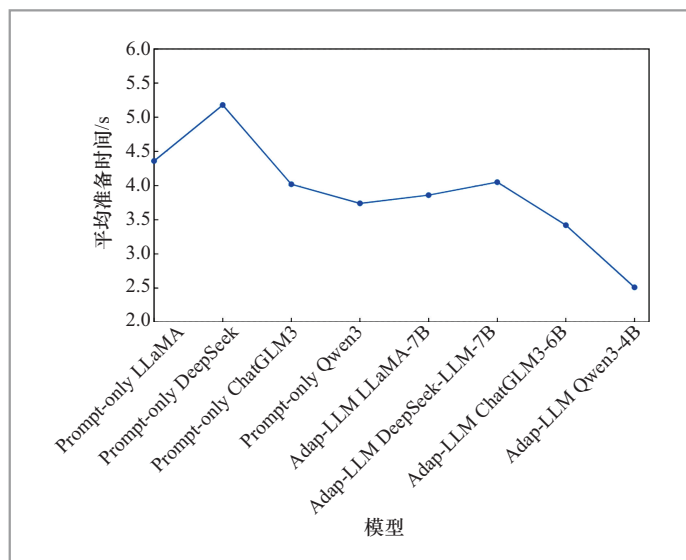


图8 不同方法下模型平均推理时间对比

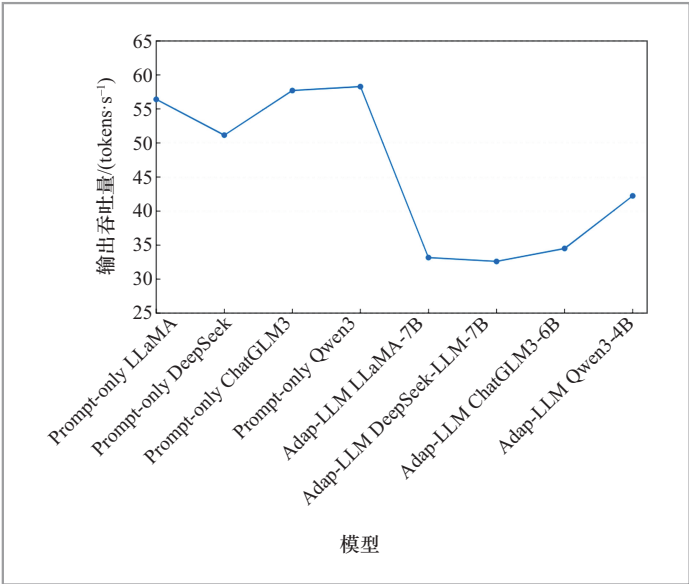


图9 不同方法下模型输出吞吐量对比

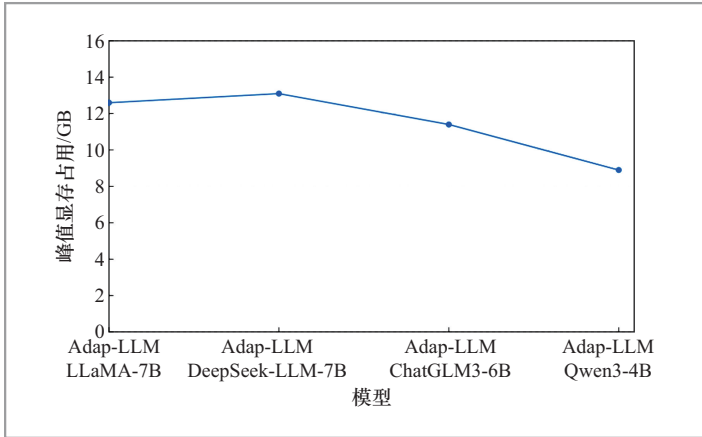


图10 Adap-LLM 基于各基座模型的峰值显存占用对比

(4) Rule Prompt w/o Reasoning: 在 LoRA 微调中加入四维评分体系和分级规则，但去除分步推理引导。

(5) Adap-LLM: 采用本文完整方法，同时引入领域知识提示词、分步推理、结构化输出约束和 LoRA 微调。

各模型的消融实验结果见表8。

由表8可知，Base 模型在未引入领域知识和参数微调的情况下表现相对有限，F1 分数在 CDE-PDS 和 PPSCD 上分别为 82.35% 和 87.91%，表明通用大语言模型虽然具备一定的语义理解能力，但仅依赖预训练知识难以充分适配电力物资供应链电子数据的专业分级需求。与 Base 相比，Knowledge Prompt 的 F1 分数在 CDE-PDS 和 PPSCD 上分别提升 2.93 个百分点和 1.92 个百分点，说明将四维评分体系、C/P/R 分级阈值和结构化输出格式显式注入提示词后，能够增强模型对分级依据和业务语义的理解能力。然而，该设置未进行参数微调，模型在面对边界样本和跨领域样本时，仍存在规则遵循不稳定的问题。LoRA w/o Knowledge 在 CDE-PDS 和 PPSCD 上分别达到 88.65% 和 92.12% 的 F1 分数，明显优于 Base 和 Knowledge Prompt，说明参数高效微调能够有效学习文档语义特征与分级标签之间的映射关

表8 消融实验结果

模型	领域知识	分步推理	结构化输出	LoRA	CDE-PDS				PPSCD			
					准确率	精确率	召回率	F1分数	准确率	精确率	召回率	F1分数
Base(基线)	×	×	×	×	83.12%	82.45%	81.56%	82.35%	88.42%	87.68%	86.95%	87.91%
Knowledge Prompt	√	×	√	×	86.27%	85.64%	84.93%	85.28%	90.73%	90.12%	89.56%	89.83%
LoRA w/o Knowledge	×	×	√	√	89.54%	88.92%	88.15%	88.65%	92.74%	91.89%	91.25%	92.12%
Rule Prompt -w/o Reasoning	√	×	√	√	91.36%	90.81%	90.24%	90.52%	94.52%	93.86%	93.41%	93.63%
Adap-LLM	√	√	√	√	92.45%	91.88%	91.67%	91.77%	95.83%	95.20%	94.61%	95.48%

系，提升模型对电子文档分级任务的适配能力。由于缺少显式领域知识引导，LoRA w/o Knowledge 模型主要依赖训练样本中的统计相关性进行判别，对法律风险、业务重要性、审计合规要求和数据敏感性等分级依据的表达仍不充分。Rule Prompt w/o Reasoning 在 LoRA 微调基础上进一步引入四维评分体系和分级规则后，其 F1 分数在 CDE-PDS 和 PPSCD 上分别达到 90.52% 和 93.63%，较 LoRA w/o Knowledge 分别提升 1.87 个百分点和 1.51 个百分点，表明领域知识与参数微调具有较好的互补作用。领域规则能够为模型提供更明确的判别边界和业务约束，但由于该设置去除了分步推理引导，模型在复杂文档和临界分数样本上仍可能出现理由不充分或分类依据不一致的问题。Adap-LLM 完整方法在两个数据集上均取得最优结果，其 F1 分数分别达到 91.77% 和 95.48%。与 LoRA w/o Knowledge 相比，Adap-LLM 在 CDE-PDS 和 PPSCD 上分别提升 3.12 个百分点和 3.36 个百分点；与 Rule Prompt w/o Reasoning 相比，分别提升 1.25 个百分点和 1.85 个百分点。实验结果表明，领域知识提示词、结构化输出约束、分步推理引导和 LoRA 微调协同作用，共同提升了模型的分级性能。其中，领域知识提示词增强了模型对分级规则和业务语义的理解，LoRA 微调提升了模型对领域数据分布的适配能力，分步推理引导则进一步提高了模型在复杂样本和边界样本上的判别稳定性。上述结果验证了本文所提 Adap-LLM 方法的有效性。

3.6 四维权重敏感性分析与边界样本性能分析

为检验四维权重配置的稳健性，本文

在 PPSCD 验证集上开展单因素敏感性分析。以法律风险程度、业务重要性、审计合规要求和数据敏感性的基准权重（0.35，0.30，0.20，0.15）为参照，每次仅扰动一个维度权重，其余维度按原始比例归一化，以保持权重和为 1。实验中固定验证样本、四维单项评分、模型推理方式及 C/P/R 分级阈值，扰动范围设为 0.10~0.50，步长为 0.05，并以 Macro-F1 为分级性能评价指标。各维度权重敏感性分析结果如图 11 所示。

由图 11 可知，各维度权重曲线在基准权重附近均保持较高水平，说明本文权重配置具有较好的稳定性。其中，法律风险程度和业务重要性对分级结果的影响相对更为显著，审计合规要求则发挥稳定的辅助判别作用；数据敏感性在权重过高时性能出现下降趋势，表明其更适合作为补充维度参与综合评分。上述结果验证了本文四维权重设置的合理性。

为进一步考察模型在临近分级阈值样本上的判别能力，本文针对表 5 中统计得到的 PPSCD 边界样本开展性能分析。与权重敏感性分析不同，边界样本分析不改变四维权重和 C/P/R 分级阈值，仅在基准权

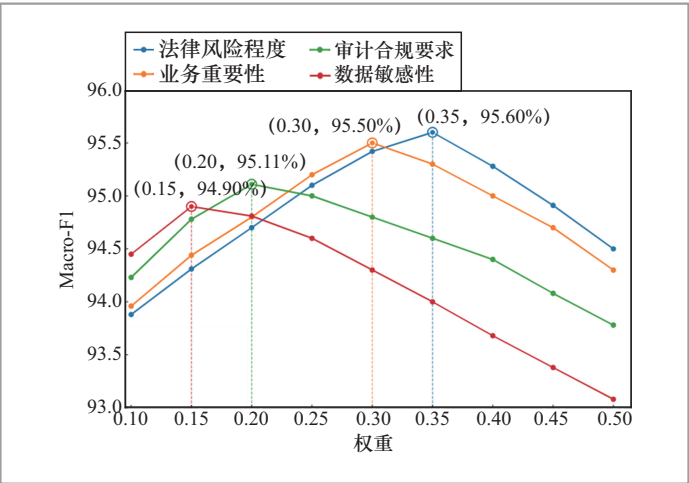


图 11 各维度权重敏感性分析结果

重配置下统计不同样本集合的准确率 (Acc) 与 Macro-F1。PPSCD 数据集边界样本性能分析如图 12 所示。

由图 12 可知，边界样本的整体性能低于非边界样本，表明临近分级阈值的样本确实具有更高的判别难度。然而，本文方法在 R/P 边界样本和 P/C 边界样本上仍保持 90% 以上的准确率和 Macro-F1，表明 Adap-LLM 方法对分级阈值附近样本仍具有较好的判别能力。

4 结束语

面向电力物资供应链电子数据信任强度分级需求，本文提出了一种融合领域知识与大语言模型的自适应分级方法——Adap-LLM。该方法基于电力物资供应链业务规则，构建了四维评分体系和 C/P/R 层次化分级标准，并通过结构化提示词与 LoRA 微调实现领域知识注入，使模型能够结合文档语义特征和专家规则完成信任强度判别。实验结果表明，Adap-LLM 方法在 PPSCD 和 CDE-PDS 两个数据集上均取得了较优性能，有效

提升了电子数据分级的准确性和稳定性。跨模型对比实验验证了 Adap-LLM 方法对不同大语言模型基座的适配能力；消融实验分别验证了领域知识注入、分步推理引导和 LoRA 微调等模块的有效性；权重敏感性分析验证了四维评分权重配置的合理性与稳健性。综合研究结果表明，将专家分级规则、结构化推理约束与参数高效微调相结合，能够为电力物资供应链电子数据的自动化、精细化和可解释分级提供有效的技术支撑。

后续研究工作将进一步引入证据片段定位、规则触发链和检索增强生成机制^[23]，以提升分级结果的可审计性和跨场景泛化能力。

参考文献：

[1] Huang Z, Liu P, Zhang Z Z, et al. Automated classification and grading of power system database fields based on rule association and probabilistic reasoning[C]//Proceedings of the 2024 10th IEEE International Conference on High Performance and Smart Computing (HPSC). Piscataway: IEEE Press, 2024: 83-88.

[2] Li Q, Li Y L, Gao J, et al. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation[C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2014: 1187-1198.

[3] Sun Y P, Qi L Z, Xu G. Classification of extra-high voltage project documents based on improved multi-label neural

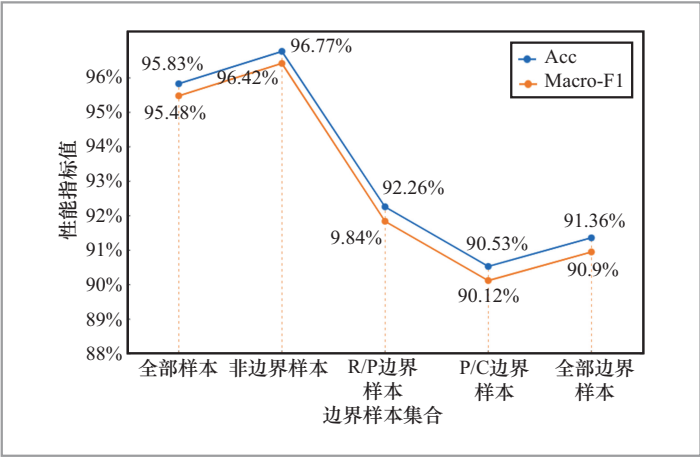


图 12 PPSCD 数据集边界样本性能分析

- network[C]//Proceedings of the 2023 5th International Conference on Applied Machine Learning (ICAML). Piscataway: IEEE Press, 2023: 202–208.
- [4] Liu T F, Hu Y L, Gao J B, et al. Hierarchical multi-granularity interaction graph convolutional network for long document classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 1762–1775.
- [5] Liu T F, Hu Y L, Wang B Y, et al. Hierarchical graph convolutional networks for structured long document classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(10): 8071–8085.
- [6] Dai S, Song D M, Wang B, et al. CCEnd-Net: cross-modal cascaded encoder-decoder network for multi-source data fusion classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 5625323.
- [7] Rajbabu K, Srinivas H, Sudha S. Industrial information extraction through multi-phase classification using ontology for unstructured documents[J]. Computers in Industry, 2018, 100: 137–147.
- [8] Zhang Y Q, Ma Z F, Luo S S, et al. Dynamic trust-based redactable blockchain supporting update and traceability[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 821–834.
- [9] Yang T G, Zhang Y, Kang C T, et al. Exploration of data classification and grading strategy for power grid enterprises[C]//Proceeding of the 2024 5th International Conference on Computer Science and Management Technology. New York: ACM, 2024: 1332–1336.
- [10] Nitti M, Girau R, Atzori L. Trustworthiness management in the social Internet of Things[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(5): 1253–1266.
- [11] Wei J, Wang X Z, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the Proceedings of the 36th International Conference on Neural Information Processing System(NeurIPS 2022). New York: Curran Associates Inc., 2022: 24824–24837.
- [12] Hu E J, Shen Y L, Wallis P, et al. LoRA: low-rank adaptation of large language models[PP]. V2. arXiv (2021–10–16)[2026–04–28].arXiv:2106.09685.
- [13] Beltagy I, Peters M E, Cohan A. Longformer: the long-document transformer [PP]. V2. arXiv (2020–12–02)[2026–04–28].arXiv: 2004.05150.
- [14] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735–1780.
- [15] Wang Z H, Yang B. Attention-based bidirectional long short-term memory networks for relation classification using knowledge distillation from BERT[C]//Proceedings of the 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCoM/CyberSciTech). Piscataway: IEEE Press, 2020: 562–568.
- [16] Cui Y M, Che W X, Liu T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Pro-

cessing, 2021, 29: 3504-3514.

[17] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models[PP]. arXiv (2023-02-27)[2026-04-28]. arXiv: 2302.13971.

[18] DeepSeek-AI. DeepSeek LLM: scaling open-source language models with longtermism[PP]. arXiv (2024-01-05)[2026-04-28]. arXiv: 2401.02954.

[19] Team GLM. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[PP]. V2. arXiv (2024-07-30)[2026-04-28]. arXiv: 2406.12793.

[20] Yang A, Li A F, Yang B S, et al. Qwen3 technical report[PP]. arXiv (2025-05-14)[2026-04-28]. arXiv: 2505.09388.

[21] 郑纬民. 构建支持大模型训练的计算机系统需要考虑的4个问题[J]. 大数据, 2024, 10(1): 1-8.

Zheng W M. Four issues to consider in building a computer system supporting large model training[J]. Big Data Research, 2024, 10(1): 1-8.

[22] 滕云, 张广艳, 孙大为, 等. 深度学习模型训练过程检查点访问性能优化方法[J]. 大数据, 2026, 12(2): 75-84.

Teng Y, Zhang G Y, Sun D W, et al. Checkpoint accessing performance optimization method for the deep learning model training process[J]. Big Data Research, 2026, 12(2): 75-84.

[23] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 9459-9474.

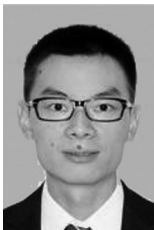
作者简介



俞春华（1973-），男，国网江苏省电力有限公司物资分公司高级经济师，主要从事供应链管理与数字化转型相关研究工作。



郭庆雷（1988-），男，博士，国网区块链科技（北京）有限公司正高级工程师，主要从事电力系统自动化与人工智能相关研究工作。



杨珂（1990-），男，博士，国网数字科技控股有限公司高级工程师，主要从事大数据分析、网络安全、人工智能相关研究工作。



柏能（1990-），男，国网江苏省电力有限公司物资分公司经济师，主要从事物资招标采购管理工作。



黄宜华（1962-），男，博士，南京大学计算机软件新技术国家重点实验室教授，主要研究方向为大模型与大数据相关理论与应用。

录用日期: 2026-06-30
通信作者: 杨珂, chenwen_crw@163.com
基金项目: 国家电网公司总部科技项目(No.5700-202418240A-1-1-ZN)
Foundation Item: State Grid Corporation of China Headquarters Science and Technology Project(No.5700-202418240A-1-1-ZN)