DOI: 10.19666/j.rlfd.202606014

融合语义消歧与知识图谱约束的发电设备多源异构数据质量提升方法

彭 强[1]，赵小红[1]，曹庆才[1]，黄言军[2]，陈子文[2]，肖祥武[2]
（1.中国大唐集团有限公司，北京 100033；
2.中国大唐集团数字科技有限公司，河北 雄安新区 071800）

［摘 要］【目的】针对发电设备多源异构数据治理中普遍存在的字段缺失、语义歧义及补全结果违反业务规则等难题，提出一种融合语义消歧与知识图谱约束的数据质量提升方法，以实现缺失字段的精准补全与故障文本的规范化。【方法】以设备标识和时间为主键，关联设备台账、SCADA 告警及运维日志构建多源语料库；基于 DeepSeek-V3 生成多类型候选补全结果，利用领域适配 BERT 模型评估语义一致性，并结合本地模块计算候选支持度；构建发电设备领域知识图谱，明确定义属性范围、设备层级、时序等强制规则与软约束；在此基础上设计双因子动态门控机制，以语义复杂度因子 S 和领域约束强度因子 C 动态融合语义评分、候选支持度及知识图谱软合规评分，通过硬约束过滤与综合重排序实现结果优选。【结果】基于某大型发电集团区域公司 2020—2024 年 76 360 条火电、风电和光伏运营数据的实验表明，该方法宏平均补全准确率、语义一致性和业务合规率分别达到 95.0%、93.9%和 95.8%，均优于单一 BERT、DeepSeek 零样本及固定权重融合等对比方法。消融实验进一步显示：去除知识图谱约束后业务合规率降至 84.3%；去除双因子门控后三项指标均出现下降，验证了各模块的有效性。【结论】所提方法通过语义理解与业务规则的深度协同，显著提升了发电设备数据补全质量与文本规范化水平，为能源电力行业的数据治理提供了可靠的技术参考。

［关 键 词］发电设备；知识图谱；动态加权；BERT；DeepSeek

［引用本文格式］彭强，赵小红，曹庆才，等. 融合语义消歧与知识图谱约束的发电设备多源异构数据质量提升方法[J]. 热力发电，年，卷(期): 1. Peng Q, Zhao X H, Cao Q C, et al. Data quality enhancement method for multi-source heterogeneous power generation equipment data integrating semantic disambiguation and knowledge graph constraints [J]. Thermal Power Generation，年，卷(期): 1.

Data quality enhancement method for multi-source heterogeneous power generation equipment data integrating semantic disambiguation and knowledge graph constraints

Peng Qiang[1], Zhao Xiaohong[1], Cao Qingcai[1], Huang Yanjun[2], Chen Ziwen[2], Xiao Xiangwu[2]
(1. China Datang Corporation Ltd., Beijing 100033, China;
2. China Datang Group Digital Technology Co., Ltd., Xiong'an 071800, China)

Abstract: [Objective] To confront the enduring challenges in governing multi-source heterogeneous data from power generation equipment-such as the frequent occurrence of missing fields, semantic ambiguity, and the persistent problem where completion results violate complex business rules-this paper introduces an innovative data quality enhancement method that seamlessly integrates semantic disambiguation with knowledge graph

收 稿 日 期：2026-06-02 修 回 日 期：2026-10-10 接 受 日 期：2026-10-12 网络首发日期：XXXX-XX-XX
基 金 项 目：中国大唐集团总部重大数字化项目（CDT0045-H-SZ-2023-15）；雄安新区重点科技项目（XA202401101005K）
Supported by：The Major Digital Project of China Datang Corporation Headquarters (CDT0045-H-SZ-2023-15); Key Science and Technology Project of Xiong'an New Area (XA202401101005K)
第一作者简介：彭强（1974），男，博士，研究员级高级工程师，主要研究方向为能源领域数字化技术，pengqiang@china-dt.com。
通信作者简介：肖祥武（1988），男，博士，高级工程师，主要研究方向为能源领域数据治理及人工智能技术应用，xiaoxw425@126.com。

constraints, aiming to achieve highly accurate missing-field completion and standardized normalization of fault-related texts. [Methods] Using equipment unique identifiers and timestamps as primary keys, a multi-source corpus is constructed by associating equipment ledgers, SCADA alarm records, and operation-maintenance logs. DeepSeek-V3 is then employed to generate multiple types of candidate completions based on existing fields and similar historical samples, while a domain-adapted BERT model-fine-tuned on power generation datasets with a partial freezing and layer-wise unfreezing strategy-evaluates the semantic consistency between candidates and contexts. The fine-tuning process uses a batch size of 16, the AdamW optimizer, and a maximum sequence length of 512. A local module further calculates candidate support degrees. Concurrently, a specialized knowledge graph for power generation equipment is built to explicitly define mandatory rules-including attribute ranges, equipment hierarchies, and temporal logic-alongside soft constraints. A dual-factor dynamic gating mechanism is designed to dynamically fuse semantic scores, candidate support degrees, and knowledge graph soft compliance scores using a semantic complexity factor S and a domain constraint strength factor C, where hard constraints strictly filter out invalid candidates and comprehensive reranking optimizes the final output. [Results] Extensive evaluations are conducted on seventy-six thousand three hundred and sixty actual operational records from thermal power, wind power, and photovoltaic systems of a regional subsidiary of a major power generation group spanning 2020 to 2024. The dataset is divided into training, validation, and test sets in a ratio of 7:1:2. The proposed method achieves macro-average completion accuracy, semantic consistency, and business compliance rates of ninety-five point zero percent, ninety-three point nine percent, and ninety-five point eight percent respectively, significantly outperforming standalone BERT, DeepSeek zero-shot, and fixed-weight fusion methods. Ablation studies further indicate that removing the knowledge graph constraints causes the business compliance rate to plummet to eighty-four point three percent, and removing the dual-factor gating degrades all three metrics, fully confirming the effectiveness of each core module. [Conclusion] By deeply synergizing semantic understanding with explicit business rules, the proposed method effectively overcomes semantic ambiguity and rule conflicts in data completion for power generation equipment, significantly improving data quality and text normalization levels, and providing a reliable technical reference for data governance in the energy and power sector.

Key words: Power generation equipment; Knowledge graph; Dynamic weighting; BERT; DeepSeek

随着能源电力系统朝着清洁化、数字化、智能化方向加速转变，发电设备的精细化运维与智能化决策对高质量数据的依赖程度日益加深[1-3]。然而，发电设备数据普遍存在语义歧义显著、字段缺失率较高、业务规则适配性欠佳等治理难题，制约了大数据与人工智能技术在设备故障诊断等方面的应用[4-6]。能源大数据治理研究已提出覆盖规划、集成、应用和维护环节的全生命周期管护框架，但面向设备缺失字段与歧义文本的候选生成、语义判别和规则一致性校验仍需协同设计[7]。随着人工智能及大语言模型（large language model, LLM）的发展，数据异常检测与智能补全成为了主流解决方案[8]。

在数据清洗与异常识别方面，文献[9-12]分别基于 Transformer 编码器、高斯过程回归、自适应 DBSCAN 和深度学习方法，面向船舶运行数据异常检测、热工系统数据驱动建模、风电机组异常数据识别和光伏数据异常修复等工业对象开展研究，为发电设备数据异常识别、建模与修复提供了方法参考。在相关数据处理研究中，文献[13-16]分别涉及不完整时序数据补全、工业大数据质量管理、电力变压器知识图谱补全和电力设备缺陷记录短文本分类，为本文的字段补全、图谱约束与语义处理提供参考。除利用传统人工智能技术外，近年来，大模型技术的应用也得到诸多研究。例如，文献[17]通过大模型和知识图谱（knowledge graph, KG）协同，构建一个跨域异质数据查询的治理方案；陈晓红等学者[18]系统分析多模态大模型对电力设备运行状态预测、电力设备故障诊断等场景的赋能作用，并提出优化数据质量的策略方法；在数据异常检测方面，文献[19]提出一种基于多模态数据的电力系统信息异常检测方法；文献[20]利用多模态数据驱动的方法开展电力模型故障诊断与预测研究。

BERT、DeepSeek 等预训练语言模型为发电设备数据治理提供新思路。BERT 基于 Transformer 编码器架构实现双向上下文理解，广泛用于自然语言处理任务，擅长深层语义理解与消歧，但生成补全长文本或数值时灵活性不足[21]；DeepSeek 大语言模型有强大问答对话、文本生成等能力，易产生“幻觉”，缺乏领域知识约束，业务合规性难保障[22]；KG 技术在设备知识融合关联方面能力较好，但传统“静态知识注入”难适配发电设备多源动态数据融合需求[23]。综上，BERT、DeepSeek 和 KG 等技术在各自擅长领域表现良好，但单一方法难以同时满足发电设备数据治理中对语义理解和业务合规

性的双重需求，复杂设备场景下的自适应融合能力有待进一步研究。

本文处理对象包括设备台账中的结构化数值、类别和日期字段，SCADA 系统中的半结构化告警记录，以及故障、缺陷和检修日志中的非结构化中文文本，不涉及原始波形、图像和音频数据。与通用文本补全任务不同，发电设备字段补全与文本规范化同时受到设备类型、运行工况、厂家参数、检修制度和安全规程等约束。本文将研究任务定义为：在给定已有字段、故障描述及领域知识的条件下，生成缺失的结构化字段值或规范化文本描述，并使结果满足语义一致性和业务合规性要求。

针对上述问题，本文提出一种基于 DeepSeek 和领域适配 BERT 模型的候选结果层外部动态评分融合与知识图谱约束协同方法。本文的核心创新与贡献主要体现在以下 3 个方面：

1）提出面向发电设备数据补全的双因子动态加权机制。基于语义复杂度和领域约束强度，对语义一致性评分、候选支持评分和知识图谱合规评分进行动态加权，提升不同缺失场景下的结果选择能力。

2）构建知识图谱约束的候选补全校验方法。围绕设备属性范围、设备—部件关系、故障—处置逻辑和检修时序规则，对大模型候选结果进行过滤和重排序，降低生成结果违反业务规则的概率。

3）建立面向发电设备数据治理的实验评价流程。针对结构化字段补全、非结构化日志规范化和规则合规校验三类任务，分别设置补全准确率、语义一致性和业务合规率等指标，并通过消融实验验证各模块作用。

1 基于动态门控机制的候选评分融合

1.1 候选评分融合机制设计

针对语义歧义与业务约束并存的发电设备数据，本文构建基于语义复杂度因子 S 和领域约束因子 C 的候选评分融合机制。在候选生成前，以设备标识为主键，将设备台账、SCADA 告警记录以及故障、缺陷和检修日志关联至相应设备实体，并结合记录时间和设备层级关系建立跨源上下文。对于每个待补全对象，从设备台账中提取设备类型、厂家参数和已有属性，从 SCADA 记录中提取运行状态及告警信息，从运维日志中提取故障现象和检修语义，再结合知识图谱中的属性范围、设备层级、时序及检修规则，构成候选生成与校验所需的统一输入。针对目标字段的数据类型，模型分别生成数值、类别、日期或规范化文本候选。DeepSeek 根据已有字段、相似历史样本和知识图谱检索规则生成候选集合；BERT 计算输入上下文与候选结果之间的语义一致性评分；本地候选评分模块计算候选支持评分；知识图谱对属性范围、设备层级、关系逻辑和时序规则进行校验。动态门控位于模型外部，根据 S 和 C 对 BERT 语义一致性评分、候选支持评分和知识图谱软约束合规评分进行归一化加权。违反强制性规则的候选先被过滤，其余候选再按综合得分重排序。整体流程包括候选生成、硬约束过滤、三类评分、动态加权与结果优选五个阶段，整体架构如图 1 所示。

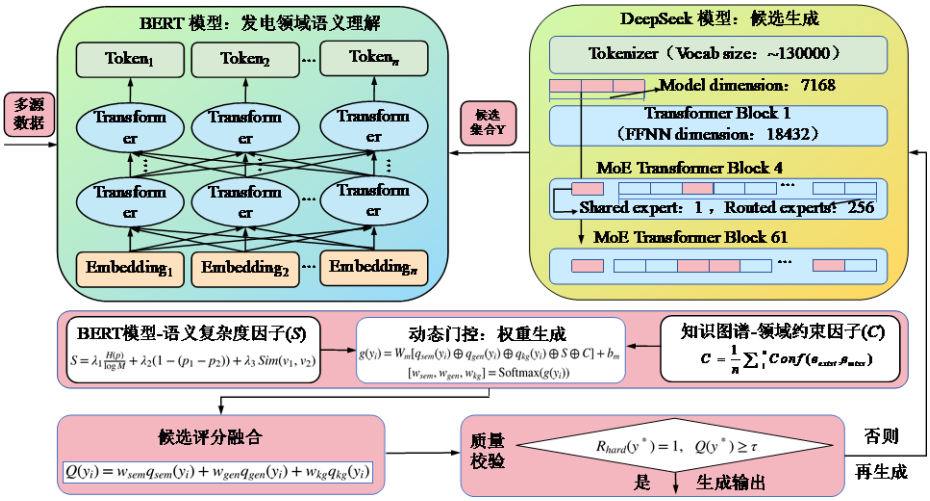


图 1 候选评分融合机制整体架构

Fig.1 Overall architecture of the candidate score fusion mechanism

1.2 特征提取

1.2.1 BERT 模型参数微调与特征提取

针对发电设备检修领域存在的风电叶片故障日志、设备检修报告等非结构化多源数据[24]，本研究以 BERT-base 为基础模型，利用发电领域数据进行参数微调，实现深层语义特征的有效提取。模型微调是提升领域适配性的关键环节。常见的微调方法包括全参数微调、分层学习率微调等[25]。结合数据集规模和领域适配需求，本文采用部分冻结—逐层解冻策略：初始阶段冻结编码器底部层，仅训练顶部层；随后按轮次逐步解冻编码器层，并依据验证集损失选择最优模型，以降低小样本条件下的过拟合风险。

BERT 编码器输出词元级上下文特征，经均值池化及特征映射得到句子级语义向量。领域微调后的特征提取过程可表示为：

$$\boldsymbol{H}_{\mathrm{BERT}}=\mathrm{Transformer}_{\mathrm{BERT}}\left(\mathrm{Embedding}(\boldsymbol{X})+\boldsymbol{PE}\right) \tag{1}$$

$$\boldsymbol{F}_{\mathrm{BERT}}=\mathrm{Pooling}\left(\boldsymbol{H}_{\mathrm{BERT}}\right)\odot\boldsymbol{W}_{\mathrm{BERT}}+\boldsymbol{b}_{\mathrm{BERT}} \tag{2}$$

式中：$\boldsymbol{X}$ 表示输入的原始文本数据；Embedding 表示融合发电领域词典的词向量嵌入层；$\boldsymbol{PE}$ 表示词元位置编码；$\boldsymbol{H}_{\mathrm{BERT}}$ 表示编码器输出的词元级特征矩阵；Pooling 表示均值池化操作；$\boldsymbol{W}_{\mathrm{BERT}}$、$\boldsymbol{b}_{\mathrm{BERT}}$ 分别表示特征映射的权重与偏置；⊙表示元素级乘法，用于特征缩放；$\boldsymbol{F}_{\mathrm{BERT}}$ 表示最终输出的句子级语义特征向量。式(1)中的“+”表示词向量嵌入与位置编码按对应元素相加，两者的词元数和隐藏维度一致。

1.2.2 DeepSeek 候选字段生成

本文调用 DeepSeek-V3 文本接口生成候选补全结果。模型输入包括设备已有字段、待补全字段名称及缺失标记、训练分区检索得到的相似历史样本和知识图谱检索规则，不包含目标字段的真实值；输出为候选补全结果集合。DeepSeek-V3 参数在实验中保持冻结，领域适配通过提示词约束、相似样本检索和后验校验实现。

针对发电设备目标字段缺失问题，本文依据训练分区中的相似样本和知识图谱规则构建提示词，由 DeepSeek 生成多个候选结果。

$$Y_{\mathrm{DS}}=\mathrm{Generate}_{\mathrm{DS}}\left(\boldsymbol{X}_{\mathrm{miss}}\mid\boldsymbol{X}_{\mathrm{exist}},K_{\mathrm{energy}}\right) \tag{3}$$

$$\boldsymbol{F}_{\mathrm{DS}}=\mathrm{Linear}\left(H_s\left(Y_{\mathrm{DS}}\right)\right)\odot\boldsymbol{W}_{\mathrm{DS}}+\boldsymbol{b}_{\mathrm{DS}} \tag{4}$$

式中：$\boldsymbol{X}_{\mathrm{miss}}$、$\boldsymbol{X}_{\mathrm{exist}}$ 和 K_{energy} 分别表示待补全字段标记、已有字段和知识图谱约束；Y_{DS} 表示由 DeepSeek 接口返回的候选集合；H_s 表示领域适配 BERT 对候选文本进行本地编码后得到的句子级表示；$\boldsymbol{W}_{\mathrm{DS}}$ 和 $\boldsymbol{b}_{\mathrm{DS}}$ 表示映射参数；$\boldsymbol{F}_{\mathrm{DS}}$ 表示式（4）的输出，该输出经归一化后形成候选支持评分。

1.3 外部动态评分单元

1.3.1 双因子驱动的动态门控机制

DeepSeek 内部的混合专家路由仅参与候选生成，本文的动态门控位于模型外部，不改变 DeepSeek 模型结构；DeepSeek 仅负责生成候选，生成完成后由各评分模块评分，再由外部门控融合评分并重排序。候选结果层的外部门控结构如图 2 所示。DeepSeek-V3 生成候选集合 Y 后，知识图谱强制规则先执行硬过滤；BERT、本地候选评分模块和知识图谱分别计算语义一致性、候选支持度和软合规性。外部门控根据语义复杂度和领域约束强度生成归一化权重，对可行候选进行综合排序。具体计算见式(5)—式(7)。

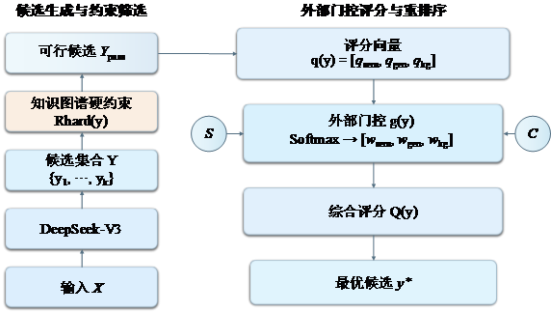


图 2 外部门控候选评分与重排序架构

Fig.2 Architecture of external-gating-based candidate scoring and reranking

$$\boldsymbol{g}\left(y_i\right)=\boldsymbol{W}_m\left[q_{\mathrm{sem}}\left(y_i\right)\oplus q_{\mathrm{gen}}\left(y_i\right)\oplus q_{\mathrm{kg}}\left(y_i\right)\oplus S\oplus C\right]+\boldsymbol{b}_m \tag{5}$$

$$w_k\left(y_i\right)=\frac{\exp\left(g_k(y_i)\right)}{\sum_{j\in\{\mathrm{sem,gen,kg}\}}\exp\left(g_j(y_i)\right)} \tag{6}$$

$$Q\left(y_i\right)=\sum_{k\in\{\mathrm{sem,gen,kg}\}}w_k\left(y_i\right)q_k\left(y_i\right) \tag{7}$$

式中：⊕仅表示拼接：将三项候选评分和 2 个因子按 q_{sem}、q_{gen}、q_{kg}、S、C 的顺序组成 5×1 列向量。$\boldsymbol{g}(y_i)$为 3×1 门控输出向量，其 3 个分量依次为 g_{sem}、g_{gen} 和 g_{kg}；$\boldsymbol{W}_{\mathrm{m}}$ 为 3×5 权重矩阵，$\boldsymbol{b}_{\mathrm{m}}$ 为 3×1 偏置向量。式（6）中，$k\in\{\mathrm{sem, gen, kg}\}$，Softmax 对同一候选的 3 个门控输出归一化，分别生成 w_{sem}、w_{gen} 和 w_{kg}，3 项权重均为非负标量且和为 1；权重随候选 y_i 变化。

式(7)中，3 项评分分别表示 BERT 语义一致性、本地候选支持度和知识图谱软约束合规性，相应权重经 Softmax 归一化；语义复杂度和领域约束强度作为门控输入，综合得分用于候选排序。语义一致

性评分由候选与上下文的 BERT 余弦相似度经线性映射得到；本地候选支持评分依据式(6)在可行候选集合内归一化；知识图谱软合规评分由候选与图谱关系及建议性规则的匹配程度计算。门控参数在训练集上学习，经验证集选择后在测试阶段冻结。

1.3.2 语义复杂度因子

语义复杂度因子 S 用于量化故障日志、运维文本等非结构化数据的歧义水平，取值范围为[0,1]，值越大表示语义越复杂、歧义越高。经发电领域微调的 BERT 模型，对输入文本进行语义标签预测，获取 Top1 和 Top2 置信度最高的语义标签，采用余弦相似度计算语义向量相似度，再通过归一化公式得到 S：

$$S=\lambda_1\cdot\frac{H(\boldsymbol{p})}{\log(M)}+\lambda_2\cdot\left(1-(p_1-p_2)\right)+\lambda_3\cdot Sim(\boldsymbol{v}_1,\boldsymbol{v}_2) \quad (8)$$

式中：H 表示语义标签分布的信息熵，M 表示候选标签数量，p_1 和 p_2 分别表示置信度最高的两个标签对应的置信度；$\boldsymbol{v}_1$ 和 $\boldsymbol{v}_2$ 表示相应的 BERT 语义向量。3 项内容分别反映预测不确定性、标签置信度接近程度和候选语义相似程度。λ_1、λ_2、λ_3 系数之和为 1，本文依次取 0.4、0.3 和 0.3，并通过验证集网格搜索确定。

1.3.3 领域知识约束因子

领域约束因子 C 用于量化目标字段与已有字段之间的规则关联强度，取值范围为[0, 1]，值越大表示候选结果受到的领域约束越强，可计算基于知识图谱中的实体关系置信度，从已有字段中抽取核心实体，查询这些实体与目标字段实体之间的关系。C 增大时，外部门控提高知识图谱软合规评分的权重。

$$C=\frac{1}{n}\sum_{i=1}^{n}\mathrm{Conf}\left(e_i,e_{\mathrm{miss}}\right) \quad (9)$$

式中：n 表示已有实体数量；Conf 表示知识图谱中实体关联的置信度，e_i 和 e_{miss} 分别表示已有实体和待补全字段对应实体；关联越强，领域约束因子越接近 1；未检索到有效关联实体时取 0。

1.4 候选评分融合

本文在候选结果层融合 3 类评分。知识图谱规则分为两类：禁止属性越界、设备层级冲突、必需处置步骤缺失和时序矛盾的规则属于强制规则，先用于过滤违规候选，得到可行集合 Y_{pass}；建议性规则、关系匹配程度和处置流程完整性用于计算软合规评分 q_{kg}。随后，外部门控依据式(10)—式(11)融合 BERT 语义评分、本地候选支持评分和 q_{kg}，从可行候选中选取综合得分最高者：

$$Y_{\mathrm{pass}}=\left\{y_i\in Y\mid R_{\mathrm{hard}}(y_i)=1\right\} \quad (10)$$

$$y^*=\arg\max Q(y_i),y_i\in Y_{\mathrm{pass}} \quad (11)$$

式中：R_{hard} 表示强制规则判定函数；Y_{pass} 表示硬约束过滤后的候选集合；y^*表示最终输出。候选只有通过全部强制规则才进入可行集合；集合为空时重新生成或转入人工审核。

1.5 质量校验

为评价方法的有效性，本文在离线测试集上采用数据补全准确率、语义一致性和业务合规率 3 个指标[26]。其中，补全准确率根据遮蔽字段的真实值计算，测试集语义一致性根据故障日志的参考文本计算，用于离线评价；在线推理使用候选语义评分、知识图谱硬约束和软合规评分进行质量控制。各评价指标定义如下：

1）数据补全准确率（$Acc_{\text{补全}}$）：用于评估发电设备结构化和半结构化记录中遮蔽目标字段的补全准确性，量化模型在数据缺失场景的修复能力。

$$Acc_{\text{补全}}=\frac{N_{\mathrm{correct}}}{N_{\mathrm{total}}}\times100\% \quad (12)$$

式中：N_{correct} 表示补全结果中与真实值一致的样本数量；N_{total} 表示参与补全的总样本数量。

2）语义一致性（$Sim_{\text{语义}}$）：校验模型输出文本与故障日志参考文本的语义一致程度，采用 Jaccard 系数与 SBERT 语义相似度的加权组合，避免单一指标的偏差。

$$Sim_{\text{语义}}=\eta\cdot\mathrm{Jaccard}\left(T_{\mathrm{gen}},T_{\mathrm{true}}\right)+(1-\eta)\cdot\mathrm{SBERT}\left(T_{\mathrm{gen}},T_{\mathrm{true}}\right) \quad (13)$$

式中：T_{gen} 和 T_{true} 分别为生成文本与参考文本；Jaccard 系数衡量词项重合程度，SBERT 余弦相似度衡量句子语义接近程度；η 为词项相似度权重，本文取 0.4。测试集语义一致性为全部文本样本得分的算术平均。

3）业务合规率（$Com_{\text{业务}}$，即冻结知识图谱规则集内部符合率）：在冻结测试分区上，按记录计算自动输出对适用规则的符合程度。规则集在测试前冻结，并按设备类型、目标字段和运行场景匹配适用规则；强制规则用于判定候选是否具备输出资格，建议性规则用于评价业务合理性。

$$Com_{\text{业务}}=\omega_1\cdot Com_{\mathrm{hard}}+\omega_2\cdot Com_{\mathrm{soft}} \quad (14)$$

式中：硬约束通过量按记录取值：自动输出通过全部适用强制规则时取 1，其余情况取 0；软约束满足度为已满足的建议性规则数占适用建议性规则总数的比例。无可行候选或无自动输出时，两项均

按 0 处理。综合业务合规率按强制规则平均符合率和建议性规则平均满足率加权计算，ω_1 和 ω_2 分别取 0.7 和 0.3，并在测试前固定。

候选只有通过全部硬约束后才可自动输出；若不存在可行候选或系统未产生输出，该记录按 0 计入，因此强制规则平均符合率并非预设为 100%。综合业务合规率同时反映自动输出覆盖和建议性规则满足程度。补全准确率和语义一致性基于参考值计算，用于离线评价；在线阶段依据候选语义评分、知识图谱软合规评分和强制规则校验结果进行质量控制。最高候选得分或软合规评分低于验证集阈值时，系统重新生成或转入人工审核。

2 知识图谱协同大模型知识生成

LLM 与 KG 的协同包括知识图谱约束候选生成和大模型辅助图谱补全，可兼顾生成灵活性与业务合规性[27]。基于上述思路，本文构建的知识图谱协同大模型设备知识生成机制如图 3 所示。

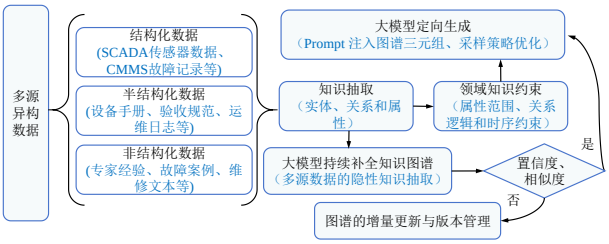


图 3 知识图谱协同大模型设备知识生成机制
Fig.3 Knowledge graph–LLM collaborative mechanism for equipment knowledge generation

注：图 3 为模块协同示意。实际入库路径为“通过置信度与相似度初筛→候选暂存区→领域专家复核→来源追溯→版本审批→增量入库及版本留痕”。未通过筛选或审核的候选不得入库；测试期间图谱版本保持冻结。

2.1 知识抽取

采用专家规则设定、DeepSeek 大模型等技术，从多源异构数据集中提取实体、关系和属性，构建发电设备领域的知识图谱。

表 1 知识图谱约束类型与示例
Tab.1 Knowledge graph constraint types and examples

约束类型	示例规则	作用
属性范围约束	逆变器额定效率应处于合理区间（如 90%～98%）	过滤数值异常
关系逻辑约束	风机振动异常应关联轴承、叶片、传动链等检查项	避免处置建议跳步
时序约束	检修日期应晚于投运日期，维护周期不得超过制度规定	保证时间逻辑
设备层级约束	部件必须属于对应设备类型	避免实体错配

本文构建的发电设备知识图谱包含实体、关系和属性，如表 1 所示列出该图谱的约束类型与示例。其中，实体包括设备、部件、故障和检修对象等，关系用于描述实体间的语义关联；属性用于记录实体或关系的特征。

2.2 领域知识的硬约束与软评分

知识图谱规则按执行方式分为硬约束和软约束。硬约束包括属性取值范围、设备—部件层级、操作先后顺序和日期逻辑等不可违反的规则，任一硬约束不通过即过滤候选；软约束包括建议检查项、关系匹配程度和处置流程完整性等，用于形成 q_{kg} 并参与候选重排序。例如，候选若违反设备温度范围或检修日期先后关系，将被直接过滤；仅缺少建议性检查项时则降低软合规评分。

2.3 知识图谱引导生成与后验校验

知识图谱分别作用于候选生成和候选校验阶段：生成前将检索到的三元组与规则注入提示词，以缩小候选空间；生成后执行硬约束过滤并计算软合规评分。本文将两类作用分别定义为生成引导和后验校验[28]：

1）生成引导：将与目标字段相关的实体、关系、属性及规则转化为 Prompt 输入 DeepSeek。

2）后验校验：对 DeepSeek 生成的多个候选执行强制规则校验，并计算候选与图谱关系及建议规则的一致程度，形成 q_{kg}；该分数用于候选重排序。

通过强制规则校验后，按式（15）计算各可行候选与知识图谱三元组的一致性。对同一输入的候选，将所得分数经极差归一化，得到知识图谱软合规评分；各候选一致性相同时，统一取中性值 0.5。

$$C_s\left(y_i,\mathrm{KG}\right)=\sum_{j=1}^{m}\left(I_{ij}^{+}-\lambda I_{ij}^{-}\right) \tag{15}$$

式中：KG 为本文构建的发电设备领域知识图谱，y_i 为第 i 个通过强制规则校验的候选；m 为参与校验的知识图谱三元组数，T_j 为第 j 条三元组；I^{+}和 I^{-}均为二值指示量，取值仅为 0 或 1。若候选 y_i 与 T_j 匹配，则 I^{+}为 1，若两者冲突，则 I^{-}取 1，其他情况下二者均为 0；λ 为冲突惩罚系数。

2.4 大模型持续补全知识图谱

通过 DeepSeek 持续补全 KG 的核心逻辑是利用大模型对非结构化数据的语义理解和隐性知识挖掘能力，填补 KG 在实体覆盖、关系链路、属性细节上的缺口，形成动态迭代机制。在发电设备领域，该过程通过以下技术路径实现。

1）多源数据的隐性知识抽取

以非结构化数据为输入，通过大模型的命名实体识别、关系抽取与属性抽取能力，挖掘潜在知识。

2）补全知识的置信度量化与筛选

为降低错误知识进入图谱的风险，候选知识的可靠性（$\mathrm{Conf}_{补全}$）由模型抽取置信度及其与图谱同类知识的关系相似度共同衡量，计算式如下。

$$\mathrm{Conf}_{补全}=\alpha_1\times P_{模型}+\alpha_2\times Sim\left(t,T_{同类}\right) \quad (16)$$

式中：t 表示由非结构化运维文本抽取的候选知识三元组，由主体实体、关系和客体实体构成；$T_{同类}$ 表示冻结知识图谱中与 t 属于同一知识类型的已确认三元组集合；$P_{模型}$ 表示大模型抽取候选三元组的置信度；Sim 表示候选三元组与图谱中同类型知识的相似度；α_1、α_2 为权重系数，满足两者和为 1，本文 α_1 和 α_2 分别取 0.6、0.4。

3）图谱的增量更新与版本管理

经置信度公式筛选的 LLM 抽取结果仅进入候选暂存区，不因 $\mathrm{Conf}_{补全}$ 达到阈值而直接写入生产知识图谱或升级为硬约束。候选知识须经领域专家复核、来源追溯和版本审批后方可入库；每次变更均保留版本号与回滚记录。实验期间用于检索、生成约束和离线评价的知识图谱版本在测试前冻结，测试样本及其输出不得反向更新该版本。

3 实例分析

3.1 数据集与模型参数设置

为验证所提方法对不同电源类型和数据形态的适用性，采用国内某大型发电集团下属某区域公司 2020—2024 年的实际运营数据，覆盖 10 台火电机组、320 台风电机组和 150 个光伏单元。数据集共 76 360 条记录，如表 2 所示，其中 9 062 条至少含 1 个目标字段缺失，缺失记录率为 12.9%；含缺失记录数是指至少 1 个目标字段缺失的记录，同一记录仅计 1 条。故障日志均关联至具体设备，并按电源类型纳入相应子集，用于文本规范化和语义一致性评价，不纳入缺失记录率统计。各类记录按 7∶1∶2 划分为训练集、验证集和测试集，对应 53 452、7 636 和 15 272 条；其中结构化/半结构化记录分别为 49 014、7 002 和 14 004 条，故障日志分别为 4 438、634 和 1 268 条。划分时先按数据来源和电源类型分层：台账记录按设备标识成组分配，保证同一台账实体不跨分区；SCADA 和日志记录在设备内按时间顺序切分，较早记录用于训练，较晚记录用于验证或测试，以避免未来信息进入训练。训练集用于构建相似样本检索库，验证集用于选取参数和阈值，测试集在模型选择前冻结。

表 2 实验数据集统计信息

Tab.2 Statistics of the experimental datasets

数据类别	数据来源	样本量	字段数	含缺失记录数	缺失记录率/%
火电设备台账	设备管理系统	12 860	36	1 415	11.0
风电 SCADA 数据	SCADA 系统	48 520	24	6 783	14.0
光伏设备台账	资产管理系统	8 640	28	864	10.0
故障日志	缺陷/工单系统	6 340	—	—	—

数据划分后，对每条结构化或半结构化记录遮蔽 1 个原本非缺失且可评价的目标字段，以原业务系统字段值为参考值；原有自然缺失字段保留为输入缺失信息；补全准确率以原本非缺失的遮蔽字段为评价对象。故障日志以原始故障描述为输入，以同一闭环缺陷或检修工单中的终审结论作为参考文本。为防止标签泄漏，测试目标值及其对应的实例级属性三元组不进入提示词、相似样本检索库或候选评分模块；知识图谱仅保留测试前冻结的型号级属性范围、实体关系和业务规则。结构化和半结构化记录按“字段名—字段值”序列化，故障日志保留自然语言形式；火电、风电和光伏训练记录均参与 BERT 微调。外部门控以每条训练记录对应的候选集合为排序单元，根据参考值生成候选正确性标签，并与 BERT 分开优化。DeepSeek 参数、知识图谱及检索库均在测试前冻结。按批次大小 16、最多 10 轮计算，BERT 微调为 33 410 个优化步，其中前 3 341 步用于学习率预热；微调参数具体见表 3。

表 3 BERT 模型微调过程关键参数

Tab.3 Key parameters for BERT model fine-tuning

序号	参数类型	设定值	说明
1	批次大小	16	依据验证集稳定性设定
2	学习率预热步数	3341 步	最大优化步数的 10%
3	最大训练轮数	10	在训练预算内选择最优检查点
4	优化器	AdamW	采用权重衰减的参数更新
5	序列最大长度	512	超长序列截断、短序列填充
6	基础模型	bert-base-chinese	中文预训练语言模型
7	训练样本数	53452 条	覆盖火电、风电和光伏训练分区
8	验证样本数	7636 条	用于参数选择，与测试集隔离
9	学习率	2×10^{-5}	依据验证集选择并在测试阶段固定
10	冻结层数	前 6 层	保留底层通用特征
11	解冻策略	第 3 轮逐层解冻	每轮解冻 2 层
12	损失函数	Cross Entropy Loss	BERT 语义标签分类

3.2 动态门控权重响应分析

外部门控对语义复杂度因子 S 和领域约束因子

C 在预设门控参数下的受控敏感性计算结果如图 4 所示。其中，保持 3 项候选基础评分分布不变：固定 C 为 0.45 考察 S 的变化，固定 S 为 0.45 考察 C 的变化；每个区间生成 200 个仿真点，并依据式(5)和式(6)计算归一化权重。结果表明，随着 S 增大，BERT 语义评分权重由 25.3%增至 75.7%；随着 C 增大，知识图谱软合规评分权重由 20.8%增至 70.8%。两图分别突出语义评分权重和知识图谱软合规评分权重的变化；图示用于检验门控函数的响应方向，其中阴影表示每个区间 200 个仿真点的 5%至 95%分位范围。

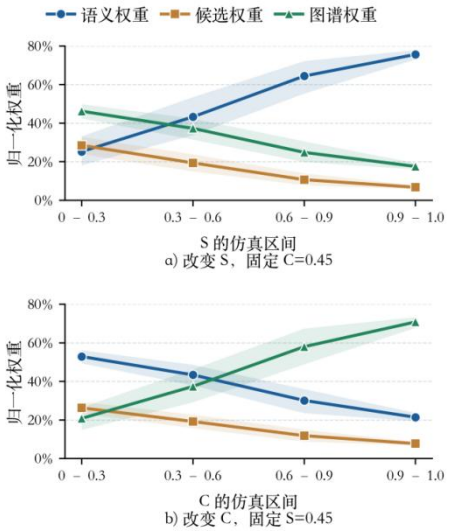


图 4 双因子驱动的外部门控权重受控仿真结果
Fig.4 Controlled simulation results of dual-factor-driven external gating weights

3.3 案例验证与结果分析

本文从冻结测试分区选取火电、风电和光伏 3 个典型记录，依次呈现目标任务、可见输入、候选生成、规则校验及最终输出。这些案例用于展示语义消歧、字段补全和规则约束的执行过程。

案例 1 针对火电汽轮机台账字段缺失和简略温度告警，检验高语义复杂度场景下的文本规范化及检修规则约束。建立如下知识图谱 5，表示其汽轮机、轴承测点、润滑系统和检修规则之间的关系。经计算该案例的语义复杂度为 0.85，属于高歧义区间；外部门控提高语义评分作用，知识图谱分别执行检修周期硬过滤和候选软合规评分，结果见表 4。

案例 1 表明，语义评分用于选择补足简略告警中关键状态信息的候选，知识图谱用于排除违反检修周期和时序要求的候选，两类机制在候选结果层分工明确。在 18 个月与 24 个月候选均满足检修间隔上限，规则 R1 只能排除 30 个月候选，不能单独确定最终补全值。18 个月候选的选取还取决于语义评分、候选支持度与软合规评分的综合排序。因此，与参考值一致的补全结果不能仅归因于硬规则过滤，还需结合合规候选之间的排序过程分析。

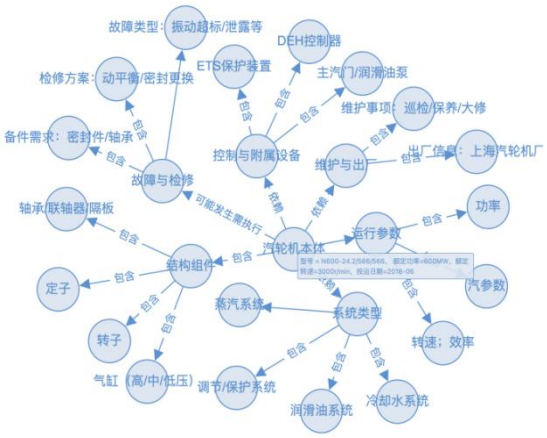


图 5 某火电厂汽轮机设备的知识图谱
Fig.5 Knowledge graph of steam turbine equipment in a thermal power plant

表 4 案例 1：某火电汽轮机本体设备数据治理
Tab.4 Case 1: Data governance for steam turbine main equipment in a thermal power plant

任务	事项	结果
数据补全	目标与可见输入	汽轮机 ID：HT-101；额定功率：600 MW；样本日期：2024-06-01；维护周期字段缺失
	候选生成	从训练分区同类型设备记录生成 18、24 和 30 个月三个候选，并调用检修周期规则 R1 校验
	输出与效果	30 个月候选违反不超过 24 个月的强制规则并被过滤；最终输出 18 个月，与测试集参考值一致
语义消歧	目标与可见输入	告警文本为“1 号轴承温度高”；可见字段包括轴承金属温度 82.6 ℃、告警阈值 80.0 ℃和机组负荷 92%
	规范化结果	1 号轴承金属温度为 82.6 ℃，高于告警阈值 2.6 ℃；建议先核查测点有效性，并结合润滑油温、油压及冷却状态排查
	案例效果	规范化文本补入了可验证的测量值和阈值信息，未在缺少证据时直接断定故障原因
业务合规性	候选	候选下次检修日期为 2026-12-01，与样本日期间隔 30 个月
	知识图谱校验	规则 R1 要求检修日期晚于样本日期，且检修间隔不超过 24 个月
	最终结果	原候选被硬过滤，系统选择 2025-12-01 的 18 个月候选；案例说明语义评分与规则过滤分别约束文本和日期结果

案例 2 针对风电机组叶片检查计划缺失和振动告警信息不足，检验日期字段补全及维护处置顺序约束，依据上述方法，建立如下知识图谱图 6，表示风电机组、振动测点、传动链、叶片及变桨系统之间的关系，图中测点与部件的关联用于支撑规则 R2 的分层排查。模型依据上次检查日期、样本日期和检查周期生成日期候选；知识图谱将未经测点与工况核查便直接调整桨距的候选判定为不合规，结果见表 5。

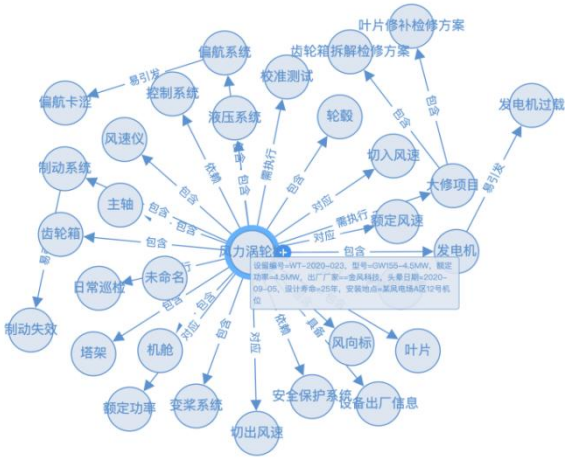


图 6 某风电场 4.5 MW 风力涡轮机的知识图谱

Fig.6 Knowledge graph of a 4.5 MW wind turbine in a wind farm

表 5 案例 2：某风电场风力涡轮机设备数据治理

Tab.5 Case 2: Data governance for wind turbine equipment in a wind farm

任务	事项	内容
数据补全	目标与可见输入	风电机组 ID：WT-2020-023；上次叶片检查日期：2023-03-15；样本日期：2024-02-20；下一次计划检查日期缺失
	候选生成	根据训练分区检查记录和高风区检查周期规则 R2 生成 2024-03-15、2024-06-15 等日期候选
	输出与效果	2024-06-15 候选超过 12 个月检查间隔并被过滤；最终输出 2024-03-15，与测试集参考值一致
语义消歧	目标与可见输入	SCADA 告警为“振动值异常”；测点 A 归一化振动指标为 1.50，告警阈值为 1.00，传感器自检状态正常
	规范化结果	测点 A 振动指标超过告警阈值，建议先核查运行工况，再检查轴承、传动链及叶片/变桨系统
	案例效果	结果保留异常事实和排查顺序，没有虚构振动单位或直接归因于单一部件
业务合规性	候选	候选维护动作是“直接调整叶片角度”
	知识图谱校验	规则 R2 要求先核查测点和运行工况，再检查轴承、传动链及叶片/变桨系统；原因未确认前不得直接调整桨距
	最终结果	不合规候选被硬过滤，系统输出分层排查建议；案例说明知识图谱能够约束维护处置顺序

案例 3 针对光伏逆变器额定转换效率缺失和组串输出异常，检验结构化数值补全及跨字段更换规则。模型从训练分区的同型号主数据生成效率候选，并结合组串电流、辐照条件、运行年限和复测要求校验维护建议，结果见表 6。图 7 表示光伏逆变器、组串、运行指标和更换规则之间的图谱关系。通过案例 3 计算表明，同型号主数据可为缺失字段提供候选，跨字段硬约束可排除缺少检测依据的更换建议。额定转换效率与归一化输出效率在校验中分别使用，不互作替代。

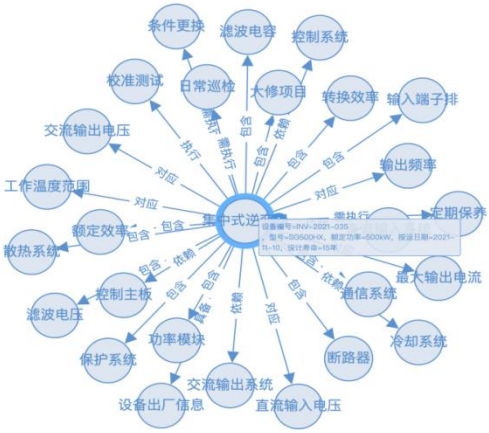


图 7 某光伏电站集中式逆变器的知识图谱

Fig.7 Knowledge graph of a centralized inverter in a photovoltaic power plant

表 6 案例 3：某光伏电站逆变器设备数据治理

Tab.6 Case 3: Data governance for inverter equipment in a photovoltaic power plant

任务	事项	内容
数据补全	目标与可见输入	逆变器 ID：INV-2021-035；型号：SG500HX；输入电压：600 V；额定功率：500 kW；额定转换效率缺失
	候选生成	从训练分区同型号主数据生成 97.2%、96.8%等候选，并用知识图谱属性范围规则排除异常值
	输出与效果	最终输出额定转换效率 97.2%，与测试集参考值一致
语义消歧	目标与可见输入	告警文本为“组串-5 输出异常”；组串电流为 8.2 A，同汇流箱其他组串电流中位数为 10.4 A，辐照度和温度处于稳定区间
	规范化结果	组串-5 电流较同汇流箱其他组串中位数低约 21.2%，建议检查遮挡、污染、接插件和绝缘状态，并结合 I-V 曲线复测
	案例效果	结果量化了异常幅度，同时避免在未完成检测时直接判定组件失效
业务合规性	候选	归一化输出效率为 88%，运行 4 年；候选维护动作是“更换组件”
	知识图谱校验	规则 R3 规定：归一化输出效率不高于 80%且复测异常，或达到规定年限并经检测不合格时，方可生成更换候选
	最终结果	更换候选被硬过滤，系统输出清洁、接插件与绝缘检查及 I-V 曲线复测建议；案例说明跨字段规则可抑制依据不足的更换建议

上述案例展示了单条记录的处理过程。图 8 进一步给出火电、风电和光伏测试子集的总体结果。火电的补全准确率、语义一致性和业务合规率分别为 95.1%、93.3%和 95.5%；风电分别为 94.7%、92.8%和 95.9%；光伏分别为 95.3%、95.6%和 96.1%。对 3 类设备等权计算得到宏平均 95.0%、93.9%和 95.8%。风电语义一致性相对较低，可能与动态工况及告警表达差异增加文本规范化难度有关。案例 1 和案例 2 表明，日期补全需满足检修或检查间隔规则；案例 3 中，额定转换效率与归一化输出效率不能混作更换依据。告警规范化还应对应可见测点、阈值和工况，避免在缺少复测结果时直接归因。

3 项指标分别对应补全、语义和合规评价。

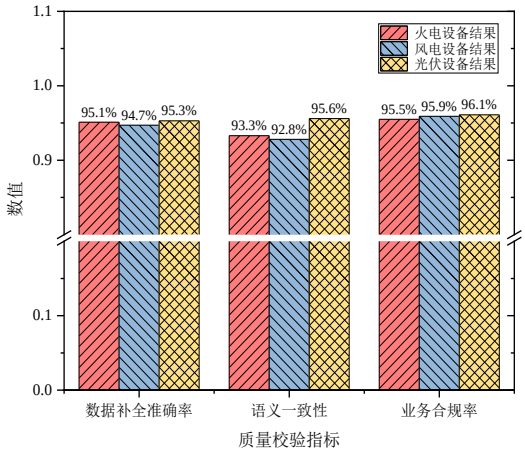


图 8 不同电源类型主要设备数据质量校验结果
Fig.8 Data quality validation results for major equipment across different power generation types

3.4 不同方法对比与消融验证

为比较候选评分融合与知识图谱约束的作用，本文设置单模型、固定权重及消融方法。各方法在同一冻结测试分区上的结果见表 7。

表 7 不同方法对比实验结果
Tab.7 Comparative experimental results of different methods

方法	补全准确率/%	语义一致性/%	业务合规率/%	平均推理时间/s
KNN 插补法	78.3	—	—	0.12
MICE 多重插补	81.6	—	—	0.35
单一 BERT 模型	86.8	90.4	83.8	0.28
DeepSeek Zero-shot	89.2	91.8	80.5	2.15
BERT+DeepSeek 固定权重	91.4	93.2	88.6	2.43
DeepSeek+KG Prompt	92.7	94.5	92.7	2.68
去除 KG 约束	93.5	95.3	84.3	2.51
去除 S、C 因子	92.1	91.2	87.5	2.48
本文方法	95.0	93.9	95.8	2.56

对比实验包括 KNN、MICE、单一 BERT、DeepSeek Zero-shot、BERT 与 DeepSeek 固定权重、DeepSeek+KG Prompt 以及两组消融方法。各方法使用同一冻结测试分区，并在各自适用任务中采用相同的目标遮蔽和参考值设置。除 KNN 和 MICE 外，补全准确率在共 14 004 条结构化/半结构化测试记录上评价，每条记录在分区完成后仅遮蔽 1 个原本非缺失的目标字段；语义一致性在共 1268 条故障日志上评价，以对应闭环工单终审结论为参考文本；业务合规率在共 15272 条测试记录上评价，按设备类型、目标字段和运行场景匹配冻结规则进行记录级统计。3 项指标先在火电、风电和光伏子集中计算，再对 3 类设备等权平均。在表 7 中 KNN 和 MICE 的算法模型仅在结构化字段补全子集上评价，不报告语义一致性和业务合规率；其准确率仅作结构化字段子集参考，不据此与 3 项宏平均作同口径优越性推断。平均推理时间为单条记录从输入到最终候选输出的端到端平均耗时，不包括离线训练和索引构建。

通过消融实验，结果显示在去除知识图谱约束后，补全准确率由 95.0%降至 93.5%，业务合规率由 95.8%降至 84.3%，而语义一致性由 93.9%升至 95.3%。该结果说明知识图谱主要改善候选的业务合规性和字段补全准确率，同时严格规则过滤与文本语义相似度之间存在权衡。去除 *S*、*C* 因子后，3 项指标均下降，说明双因子门控对不同场景的评分组合具有调节作用。

与单一模型和固定权重方法相比，本文方法 3 项宏平均指标的点估计均取得更高结果；知识图谱消融的语义一致性变化反映上述权衡，不据此作全面优越性判断。

4 总结与讨论

本文提出一种面向发电设备结构化、半结构化与文本数据的数据质量提升方法。DeepSeek 负责生成候选，BERT 计算语义一致性评分，知识图谱通过硬约束过滤和软合规评分参与候选优选；双因子动态门控完成候选结果层的自适应评分融合。对某发电企业 2020-2024 年运营数据分析，火电、风电和光伏 3 类设备的宏平均补全准确率、语义一致性和业务合规率分别为 95.0%、93.9%和 95.8%；与单模型和固定权重方法相比，3 项宏平均指标的点估计均更高。消融实验中，去除知识图谱约束后业务合规率下降，去除 *S*、*C* 因子后 3 项指标均下降，表明两个模块分别影响规则合规性和多场景评分融合。未来可基于本文研究成果，重点探索以下方面：1）进一步优化动态加权机制，引入更多领域特征因子，提升模型在复杂场景下的自适应能力；2）扩展知识图谱的覆盖范围，增强对多样化设备类型的支持能力；3）探索模型在实时性要求更高的工业场景中的应用潜力；4）研究跨电厂、跨区域的数据治理迁移学习方法，提升方法的泛化能力。

[参 考 文 献]

[1] 肖祥武, 王丰, 王晓辉, 等. 面向工业互联网的智慧电厂仿生体系架构及信息物理系统[J]. 电工技术学报, 2020, 35(23): 4898-4911.
Xiao X W, Wang F, Wang X H, et al. Bionic structure and cyber-physical system for intelligent power plant oriented

to the industrial Internet [J]. Transactions of China Electrotechnical Society, 2020, 35(23): 4898-4911.

[2] 曾鸣, 许彦斌, 王晟嫣, 等. 基于数据管护理论的能源互联网数据资产治理方法及关键问题研究[J]. 电网技术, 2020, 44(7): 2420-2429.

[3] 廖美英, 胡列豪, 张勇军. 人工智能技术在发电厂智能化建设中的应用与挑战[J]. 广东电力, 2024, 37(11): 109-119.
Liao M, Hu L, Zhang Y. Applications and challenges of AI technology in power plant intelligent construction[J]. Guangdong Electric Power, 2024, 37(11): 109-119.

[4] 王臻, 刘东, 徐重酉, 等. 新型电力系统多源异构数据融合技术研究现状及展望[J]. 中国电力, 2023, 56(4): 1-15.
Wang Z, Liu D, Xu C, et al. Status quo and prospect of multi-source heterogeneous data fusion technology for new power system[J]. Electric Power, 2023, 56(4): 1-15.

[5] 顾菊平, 赵佳皓, 张新松. 电力设备多参量监测数据清洗研究现状及展望[J]. 高电压技术, 2024, 50(8): 3403-3420.
Gu J P, Zhao J H, Zhang X S. Research review and prospect of data cleaning for multi-parameter monitoring data of power equipment[J]. High Voltage Engineering, 2024, 50(8): 3403-3420.

[6] 闫佳和, 李红辉, 马英, 等. 多源异构数据融合关键技术与政务大数据治理体系[J]. 计算机科学, 2024, 51(2): 1-14.
Yan J H, Li H H, Ma Y, et al. Multi-source Heterogeneous Data Fusion Technologies and Government Big Data Governance System[J]. Computer Science, 2024, 51(2): 1-14.

[7] 刘露, 杨晓雷. 新基建背景下的数据治理体系研究——以数据生命周期为总线的治理[J]. 治理研究, 2020, 36(4): 59.

[8] 冯钰瑶, 张宏云, 王鹏飞, 等. 推动全生命周期数据治理的关键技术选择与对策[J]. 中国科学院院刊, 2025, 40(10): 1696-1708.
FENG Y, ZHANG H, WANG P, et al. Key technology selection and countermeasures to promote full lifecycle data governance[J]. Bulletin of Chinese Academy of Sciences, 2025, 40(10): 1696-1708.

[9] Liu H, Wu C, Li B, et al. Research on ship anomaly detection algorithm based on transformer-GSA encoder[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(6): 8752-8763.

[10] 贾昊. 基于高斯过程回归的热工系统数据驱动建模研究[D]. 华北电力大学 (北京), 2021.

[11] 雷萌, 郭鹏, 刘博嵩. 基于自适应 DBSCAN 算法的风电机组异常数据识别研究[J]. 动力工程学报, 2021, 41(10): 859-865.
Lei M, Guo P, Liu B S. Study on abnormal data recognition of wind turbines based on adaptive DBSCAN algorithm[J]. Journal of Chinese Society of Power Engineering, 2021, 41(10): 859-865.

[12] 薛文琦. 基于深度学习的光伏数据异常修复与出力预测方法研究[D]. 湖南大学, 2021.

[13] 周立, 吴洋洋, 苗晓晔. 基于深度学习的不完整时序数据补全方法综述[J]. 计算机研究与发展, 2025, 62(9): 2233-2258.

[14] 唐军. 工业大数据质量管理关键技术研究及应用[D]. 四川大学, 2024.

[15] 徐浩, 康振渊, 张焱, 等. 面向电力变压器故障辅助决策的知识图谱构建及补全策略研究[J]. 中国安全生产科学技术, 2025, 21(5): 46-54.
Xu H, Kang Z Y, Zhang Y, et al. Research on construction and completion strategies of knowledge graph for fault assistant decision-making of power transformers[J]. J Saf Sci Tech, 2025, 21(5): 46-54.

[16] Ye Z, Guo B, Qi D. BERT‐TriF: An inductive short text classification model for power equipment defect records[J]. Engineering Reports, 2023, 5(10): e12671.

[17] 吴文隆, 尹海莲, 王宁, 等. 大语言模型和知识图谱协同的跨域异质数据查询框架[J]. 计算机研究与发展, 2025, 62(3): 605-619.

[18] 陈晓红, 傅文润, 刘朝明, 等. 人工智能大模型在电力设备运维场景中的应用探讨[J]. 中国工程科学, 2025, 27(1): 180-192.
Chen X, Fu W, Liu C, et al. Application of artificial intelligence large language model in power equipment operation and maintenance[J]. Strategic Study of Chinese Academy of Engineering, 2025, 27(1): 180-192.

[19] Chen L, Zhou X X, Zhou P, et al. Anomaly detection method for power system information based on multimodal data[J]. PeerJ Computer Science, 2025, 11: e2976.

[20] He Y, Jiang W, Zhao H, et al. Fault Diagnosis and Prediction of Large Power Model Driven by Multi-Modal Data[C]//2024 11th International Forum on Electrical Engineering and Automation (IFEEA). IEEE, 2024: 66-71.

[21] 张智源, 孙水华, 徐诗傲, 等. 基于 BERT 和多窗口门控 CNN 的电机领域命名实体识别[J]. 计算机应用研究, 2023, 40(1): 107-114.
Zhang Z, Sun S, Xu S, et al. Named entity recognition in motor field based on BERT and multi-window gated CNN[J]. Appl. Res. Comput, 2023, 40(1): 107-114.

[22] 武延军. DeepSeek 引发的 AI 创新和开源生态发展的思考[J]. 中国科学院院刊, 2025, 40(3): 446-452.
WU Y. Thoughts on AI innovation and open source development: Lessons from DeepSeek[J]. Bulletin of Chinese Academy of Sciences, 2025, 40(3): 446-452.

[23] 李晓璐, 王岢, 赵冰, 等. 面向风电机组运维数据的知识图谱构建方法[J]. 动力工程学报, 2024, 44(6): 886-894.

[24] 王志敏, 黄骞, 王可轩, 等. 智慧电厂大数据云平台的架构建设与模型开发研究[J]. 动力工程学报, 2025, 45(2): 282-291.
Wang Z M, Huang Q, Wang K X, et al. Architecture Construction and Model Development of Big-data Cloud Platform for Intelligent Coal-fired Power Plants[J]. Journal of Chinese Society of Power Engineering, 2025, 45(2): 282-291.

[25] 吴春志, 赵玉龙, 刘鑫, 等. 大语言模型微调方法研究综述[J]. 中文信息学报, 2025, 39(2): 1-26.

[26] 黄倩倩, 赵正, 刘钊因. 数据流通交易场景下数据质量综合管理体系与技术框架研究[J]. 数据分析与知识发现, 2022, 6(1): 22-34.
Huang Q, Zhao Z, Liu Z. Comprehensive management system and technical framework of data quality in the data circulation transaction scenario[J]. Data Analysis and Knowledge Discovery, 2022, 6(1): 22-34.

[27] 郑强, 许振彬. 面向特种设备的大语言模型-知识图谱双向推理优化与幻觉抑制方法[J]. 数据采集与处理, 2025, 40(3): 647-658.

[28] Yang J, Lu G, He S, et al. A novel model for relation prediction in knowledge graphs exploiting semantic and structural feature integration[J]. Scientific Reports, 2024, 14(1): 12962.

（责任编辑 XXX）