ICS 国际标准分类号
CCS 中国标准文献分类号

# 团　　体　　标　　准

T/CES XXX-XXXX

# 电力知识智能检索流程规范

Specification for intelligent retrieval process of electric power knowledge

（征求意见稿）

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会　发布

# 目　　次

# 前 言

本文件按照 GB/T1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》给出的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本文件由中国电工技术学会提出。

本文件由中国电工技术学会标准工作委员会源智慧化工作组归口

本文件起草单位：国网信息通信产业集团有限公司、国家电网有限公司大数据中心、中国电力科学研究院有限公司、国网智能电网研究院有限公司、北京国网信通埃森哲信息技术有限公司、四川中电启明星信息技术有限公司、国网福建省电力有限公司

本文件主要起草人：李强、赵峰、刘迪、邱镇、陈振宇、李博、刘识、李炳森、黄晓光、王晓东、张琳瑜、秦余、张国梁、邹达明、商峻、郭厅、李文璞、赵浩东、张桉恺、宋卫平、杨帆、高攀、王红蕾、董梅、李欢欢、徐小云、叶林峰、赵林林、王誉博、李扬笛、谢炜、林爽

本文件为首次发布。

# 电力知识智能检索流程规范

## 1 范围

本文件规定了对电力知识智能检索流程的电力知识文本基本要求、知识库构建要求和电力知识检索流程要求，其中电力知识智能检索流程主要针对电力行业文本数据。

本文件适用于国内各单位实现电力知识智能检索流程的相关人员。

## 2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 5271.28 信息技术 词汇 第28部分:人工智能 基本概念与专家系统

T/CESA 1040-2019 信息技术 人工智能 面向机器学习的数据标注规程

T/CES 128-2022 电力人工智能平台总体架构及技术要求

T/CES 156-2022 电力智能交互文本训练语料标注规范

## 3 术语和定义

下列术语和定义适用于本文件。

3.1

**电力知识 electricity knowledge**

指与电能的产生、传输和利用过程中所涉及的知识和技术。

3.2

**知识库 knowledge base**

是知识工程中结构化、易操作、易利用、全面有组织的知识集群。

3.3

**文本向量化 text vectorization**

指将文本信息表示成能够表达文本语义的向量，即用数值向量来表示文本的语义。

3.4

**向量相似度 vector similarity**

指衡量两个向量在数值上的接近程度的度量。

3.5

**智能检索 intelligent retrieval**

指计算机根据用户的检索词和检索要求，运用人工智能技术自动扩展检索词和构造检索式，以满足用户检索要求的过程。

3.6

**提示模版 prompt template**

指在AI大模型训练或应用过程中，用来引导模型生成特定类型文本或解决特定任务的一种预设语句。通过给模型提供明确的上下文信息或者参数信息，提示模板可以有效地提高模型在特定任务上的表现。

3.7

**提示 prompt**

在AI大模型中，prompt主要是用来给模型提供提示输入信息的上下文和输入模型的参数信息。

## 4 缩略语

下列缩略语适用于本文件。

LLM: 大语言模型(Large Language Model)，也称大型语言模型。

## 5 总则

本文件对电力知识智能检索流程的规范主要体现在三个方面: 电力知识文本基本要求、电力知识库构建基本要求和电力知识检索流程基本要求。其中，电力知识文本基本要求用于规范电力知识文本文件的命名、存储格式等，电力知识库构建基本要求用于规范构建电力知识库，电力知识检索流程基本要求用于规范基于电力知识库和电力大模型的电力知识智能检索流程。这三方面内容的具体组织框架如图 1 所示:

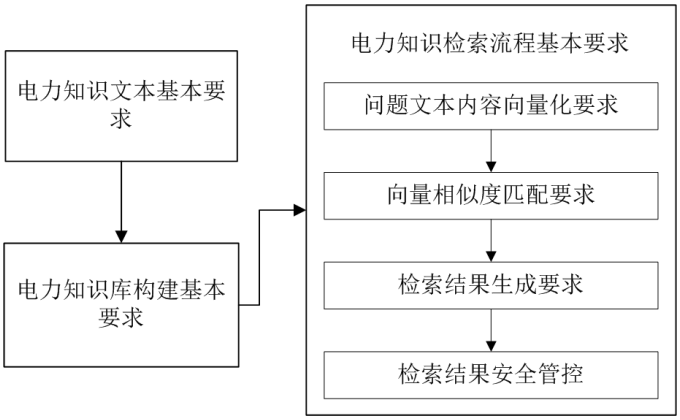


图 1　内容组织框架

## 6 电力知识文本基本要求

### 6.1 存储格式要求

电力知识文件应采用 txt、docx、md、pdf 存储格式。其中，对于 excel 等其他格式的原始数据可以转换为上述存储格式之一且应优先转成 txt 存储格式。

### 6.2 命名要求

电力知识文本文件名称应由三个部分组成：

a) 第一部分为当前电力知识文本的专业信息；

b) 第二部分为原始源文件生成时的日期，日期格式：YYYY-MM-DD；

c) 第三部分为文件唯一性编号，从 1 开始计数；

d) 这三部分用下划线连接，且文件名称长度和扩展名在内最大长度不超过 100 个字符(包含中英文字符和特殊字符)；

e) 文件命名举例：输电线路金具部件介绍_2023-10-11_1。

### 6.3 质量要求

电力知识文本文件的质量要求如下。

a）确保一个文件中的知识都是属于同一个细分领域、同一权限等级，且不同细分领域、不同权限的知识不能混杂在同一个文件中；

b) 电力知识数据中不应存在重复的记录或重复的信息，以确保数据的唯一性；

c) 电力知识数据中不应包含特殊字符、停用词、HTML 标签等；

d）电力知识数据中不应包含含糊不清、模棱两可、参考价值小、意义不大的知识内容；

e）电力知识数据中不应包含图片、表格等数据，但可将图片、表格中的内容提炼为文字表述且可优先转化成问答形式；

f）电力知识数据中包含的问答形式的数据应提供详细和全面的答案，并确保回答符合专业要求和语言规范。

### 6.4 电力知识文本描述文件要求

每批次电力知识文本文件应有一个描述文件，且描述文件应满足下述要求：

a) 存储格式应为 txt 格式；

b) 命名应由两个部分组成：

1) 本文件创建的日期，日期格式：YYYY-MM-DD；

2) 文件唯一性编号，从 1 开始计数；

3) 文件名的各部分用下划线连接，文件命名示例：2023-10-12_1。

c) 文件内容应描述本电力知识信息的基本信息，应包括电力知识文本文件的来源、创建日期、联系人、文本用途等信息。

### 6.5 安全管控

电力知识文件存储环境应满足安全管控要求。具体要求包括：

a) 电力知识文件应存储在指定安全机器中，同时该机器应开启防火墙，安装杀毒软件，并禁用 USB 接口功能；

b) 存储电力知识文件的机器中的所有数据文件需定期做好数据备份，不得擅自拷贝、传输，防止数据丢失或泄露。

## 7 电力知识库构建基本要求

### 7.1 文本内容拆分要求

电力知识文本内容的拆分方法应采用规则拆分或语义拆分，具体要求如下：

a）利用规则进行文本拆分应根据文本内容中常见终止符号进行拆分，且拆分后的文本长度控制在1000内。常见文本终止符号如：单字符断句符、中英文省略号、双引号等。

b) 利用语义拆分方法应将文本内容拆分为具有语义信息的最小块，一般是有意义的句子，再将这些小块组合并成一定大小的文本段且文本段大小控制在1000内。

### 7.2 文本向量化方法选择要求

对拆分的文本内容进行向量化时选择的向量化方法要求如下：

a）应选择可以对中文文本进行向量化的方法；

b）所选向量化方法在统一评测标准中具有较好评测结果；

c）所选向量化方法在应用过程中易调用、易迁移部署。

### 7.3 向量数据库选择要求

向量化应存储向量数据库中，选择向量数据库的具体要求如下：

a）向量数据库应具备较好查询性能，可结合向量数据库的索引技术、数据结构、硬件配置等因素判断向量数据库的性能；

b）选择的向量数据库的存储量应满足电力知识数据的存储需求；

c）所选向量数据库应具备良好的社区支持，以便更容易地解决问题和获取帮助。

d）根据业务需求选择具备不同功能的向量数据库，如部分向量数据库提供了全文搜索功能，而另一部分则提供了更专业的向量搜索功能。

## 8 电力知识检索流程基本要求

### 8.1 总体要求

电力知识智能检索流程基本要求具体包括问题文本内容向量化要求、向量相似度匹配要求、匹配结果处理要求、检索结果生成要求等，如图2所示：

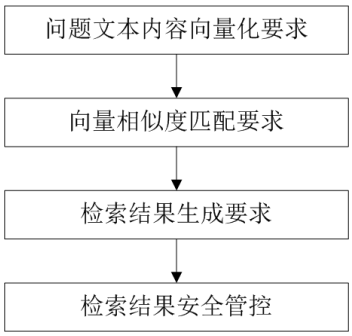


图 2 电力知识检索流程基本要求

### 8.2 问题文本内容向量化要求

对用户输入的问题文本内容进行向量化操作的基本要求具体如下：

a）对用户输入的问题文本内容进行向量化时选择文本向量化方法的要求应与第7章中文本向量化要求相同；

b）选用的文本向量化方法应与第7章选用的文本向量化方法相同；

c）用户输入的问题文本内容向量化后应根据电力业务需求，与未向量化之前的问题文本一同写入指定日志文件中。

### 8.3 向量相似度匹配要求

进行向量相似度匹配的具体要求如下：

a）应选择合适的向量匹配算法，所选向量匹配算法能够准确快速计算出问题文本向量与电力知识库中不同文本向量之间的相似度值；

b）对计算所得的相似度值按照从大到小（或者从小到大）的顺序进行排序，并选取前k个相似度值对应的文本向量作为匹配结果。其中确定k值的要求如下：

1）一般情况下，k值默认设定为3；

2）根据电力知识库向量匹配情况，可增大或减小k值以获取满足业务需求的结果。

c)将获取的k个文本向量转换为对应的文本内容，并将文本内容按照一定方式组合成一个文本段落，其中组合方式要求如下：

1）k个文本内容按照向量之间的匹配度（由大到小或者由小到大）直接组合成一个文本段落；

2）k个文本内容打乱顺序后随机排序组合成一个文本段落。

### 8.3 检索结果生成要求

检索结果生成应按照如下要求：

a）最终的检索结果通过电力语言大模型生成；

b）电力语言大模型应由基座大模型微调得到，其中基座大模型的选则应遵循以下要求：

1）基座模型LLM应可以生成中文文本内容；

2）需根据现有硬件条件(如显卡GPU的性能、数量、服务器数量等)选择基座模型，且所选基座模型LLM能够在现有硬件环境中运行；

3）基座模型LLM应易于迁移部署，且其生成文本内容的反应时间应小于电力知识智能检索要求的最长反应时间；

4）基座模型可以通过提示信息及用户指令产生指定内容；

c）电力大模型生成检索结果应根据匹配到的电力知识内容和问题内容得到。

### 8.4 检索结果安全管控

检索结果需进行安全管控，具体要求如下：

a）检索结果应以指定形式返回，如以字典形式返回:{“检索结果”：“电力知识检索具体内容”}

b）检索结果和问题内容应写入指定日志文件；

c）针对电力行业内部人员，知识检索结果一般在电力行业内部软件上返回给需求人员；

d）针对非电力行业内部人员，检索结果需对检索结果脱敏后返回给需求人员。

# 附录 A

## 资料性附录

## 电力知识智能检索提示模版应用示例

**应用场景**：电力知识智能检索

**步骤一**：用户输入问题，具体问题（question）具体为: 根据绝缘子的制成材料分类，绝缘子的类型有哪些？

**步骤二**：从电力知识库匹配相关文本内容并进行文本内容拼接，匹配到的具体内容（context）为：绝缘子按安装方式不同，可分为悬式绝缘子和支柱绝缘子；按照使用的绝缘材料的不同，可分为瓷绝缘子、玻璃绝缘子和复合绝缘子（也称合成绝缘子）；按照使用电压等级不同，可分为低压绝缘子和高压绝缘子；按照使用的环境条件的不同，派生出污秽地区使用的耐污绝缘子；按照使用电压种类不同，派生出直流绝缘子；尚有各种特殊用途的绝缘子，如绝缘横担、半导体釉绝缘子和配电用的拉紧绝缘子、线轴绝缘子和布线绝缘子等。此外，按照绝缘件击穿可能性不同，又可分为A型即不可击穿型绝缘子和B型即可击穿型绝缘子两类。

**步骤三**：已知提示模版是：prompt_template = """已知信息：{context}。根据上述已知信息，简洁和专业的来回答用户的问题。优先用已知信息的原文回答，不要解释信息来源。如果无法从中得到答案，请说“根据已知信息无法回答该问题”或“没有提供足够的相关信息”，不允许在答案中添加编造成分，答案请使用中文。问题是：{question}"""，将用户输入的问题（question）和从电力知识库中匹配并进行拼接而成的相关电力知识文本（context）两部分嵌入到该提示模版中，则得到新的提示具体如下：

prompt= """已知信息：绝缘子按安装方式不同，可分为悬式绝缘子和支柱绝缘子；按照使用的绝缘材料的不同，可分为瓷绝缘子、玻璃绝缘子和复合绝缘子（也称合成绝缘子）；按照使用电压等级不同，可分为低压绝缘子和高压绝缘子；按照使用的环境条件的不同，派生出污秽地区使用的耐污绝缘子；按照使用电压种类不同，派生出直流绝缘子；尚有各种特殊用途的绝缘子，如绝缘横担、半导体釉绝缘子和配电用的拉紧绝缘子、线轴绝缘子和布线绝缘子等。此外，按照绝缘件击穿可能性不同，又可分为A型即不可击穿型绝缘子和B型即可击穿型绝缘子两类。根据上述已知信息，简洁和专业的来回答用户的问题。优先用已知信息的原文回答，不要解释信息来源。如果无法从中得到答案，请说“根据已知信息无法回答该问题”或“没有提供足够的相关信息”，不允许在答案中添加编造成分，答案请使用中文。问题是：按照绝缘子的制成材料分类，绝缘子的类型有哪些？"""。

**步骤四**：将后新得到的提示输入到LLM模型中，即可生成问题的相应回答，且在前端展示的示例如下：

根据绝缘子的制成材料不同，可以将绝缘子的类型分为瓷绝缘子、玻璃绝缘子和复合绝缘子。