ICS 35.240.50
CCS L 77

团　　体　　标　　准

T/CES XXX-2024

基于大模型的电力企业向量知识库及增强检索应用技术框架

A technical framework for the vectorized knowledge base and enhanced retrieval application in power enterprises based on large models

20XX-XX-XX 发布　　20XX-XX-XX 实施

中国电工技术学会　发布

目　　次

前　言

本标准按照 GB/T 1.1—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》的规则起草。

请注意本标准的某些内容可能涉及专利，本标准的发布机构不承担识别专利的责任。

本标准由国网信息通信产业集团有限公司提出并解释。

本标准由中国电工技术学会标准工作委员会电力信息化标准专业委员会工作组归口。

本标准起草单位：国网信息通信产业集团有限公司、福建亿榕信息技术有限公司。

本标准主要起草人：庄莉、梁懿、王秋琳、宋立华、郑耀松、丘志强、李建华、李年勇、邢国用、张晓东、陈江海、王燕蓉。

本标准为首次发布。

本标准在执行过程中的意见或建议反馈至中国电工技术学会标准工作委员会能源智慧化工作组。

基于大语言模型的企业向量知识库构建及增强检索应用技术框架

1 范围

本标准规定了基于大模型的企业向量知识库构建和增强检索应用的技术框架和要求。

本标准适用于国内电力企业为满足企业生产经营活动智能化需求开展的基于大模型的企业向量知识库构建及增强检索应用工作。

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 41867—2022 人工智能术语

GB/T 42755—2023 人工智能面向机器学习的数据标注规程

GB/T 43782—2024 人工智能 机器学习系统技术要求

GB/T 5271.28—2001 信息技术 词汇 第 28 部分:人工智能 基本概念与专家系统》

GB/T 5271.34—2006 信息技术.词汇.第 34 部分:人工智能. 神经网络》

T/CES 156—2022 电力智能交互文本训练语料标注规范

T/CES 129—2022 电力人工智能平台样本规范

3 术语和定义

下列术语和定义适用于本文件。

3.1 大模型 large model

指代一种大语言模型，由具有许多参数（通常数十亿个权重或更多）的人工神经网络组成，使用自监督学习或半监督学习对大量未标记文本进行训练。

3.2 训练 training

教会神经网络在输入值的样本和正确输出值之间做出结合的步骤。

3.3 推理 inference

从已知前提导出结论的方法。

注 1：在人工智能领域，前提是事实或者规则。

注 2：术语“推理”既指过程也指结果。

3.4 接口 interface

两个功能单元共享的边界，它由各种特征（如功能、物理互联、信号交互等）来定义。

3.5 向量知识库 vector knowledge database

一种是专门用来存储、查询、管理向量化知识的数据库，其存储的向量化知识来自于对企业所属文本、语音、图像、视频等的向量化。

3.6 安全与隐私保护 security and privacy protection

基于大语言模型的企业向量知识库的构建、应用和管理过程中保护数据安全和用户隐私的措施和机制。

4 符号、代号和缩略语

下列符号、代号和缩略语适用于本文件。

token：关键字、标识符、运算符、分隔符、字符串字面量

5 企业向量知识库构建

5.1 技术框架

基于《GB/T 43782-2024 人工智能 机器学习系统技术要求》《T/CES 129-2022 电力人工智能平台样本规范》《GB/T 31171-2023 人工智能数据管理规范》《GB/T 31169-2023 人工智能安全评估指南》具体要求，提出基于大语言模型的企业知识向量库构建技术框架主要包括：知识数据归集与预处理、向量模型训练、知识向量化、向量知识库构建、向量知识维护与更新 5 个技术环节，如图 1 所示。

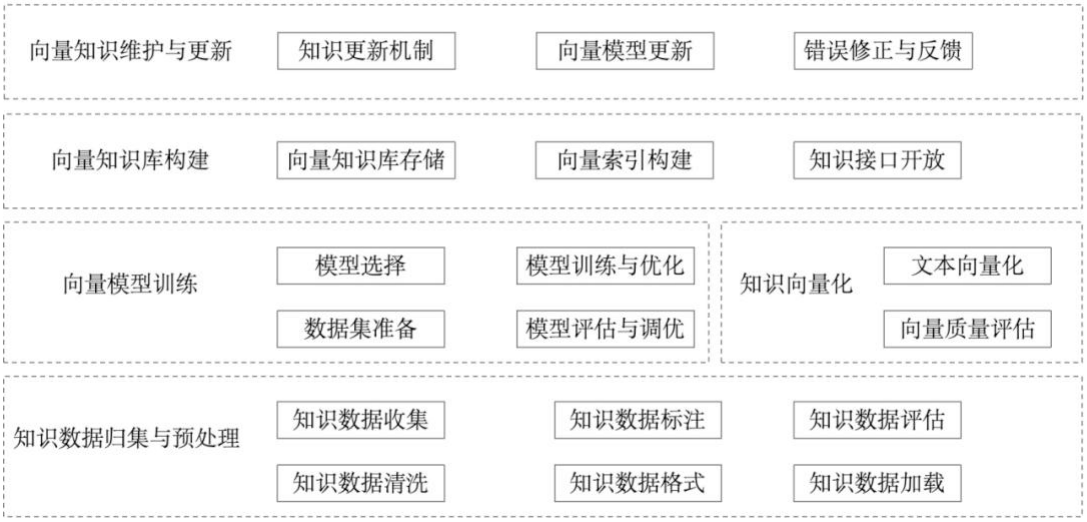

图 1　企业向量知识库构建技术框架

5.2 知识数据归集与预处理

知识数据归集与预处理包括但不限于知识数据收集、知识数据清理、知识数据标注、知识数据格式、知识数据评估、知识数据加载，具体要求如下：

a）确定数据收集范围，了解数据结构、数据量和更新频率等相关要求，以确保数据的可用性和时效性，数据范围包括但不限于企业内部的数据库、文档管理系统、业务系统及外部的公开数据集、行业报告等；

b）实施数据采集，定义关键数据指标，如数据准确性、完整性、一致性、时效性等，实施异常检测算法，识别数据异常、缺失或不一致的情况，建立自动化数据采集手段，定期检查数据源和数据流程，确保数据按计划采集、传输和处理，同时确保数据完整性与安全性，避免数据丢失或被非法获取；

c）定义数据清洗标准，实施数据进行清洗，使用去重、脱敏、大小写转换等数据清洗手段，去除重复、错误或无关的数据，数据重复率、错误率均要求低于 1%。

d）实施知识数据格式化、标准化和归一化，使其符合后续处理和分析的要求。

e）定义知识数据标注标准规范，使用标准知识格式，实施标注活动。

f）定义知识数据评估方法、测试数据集、自动化工具，应考虑从任务类别、需求与数据平衡性、领域复杂度等维度进行综合分析，在随机选取的测试集上，数据准确性要求达到 99%，数据完整性要求达到 98%，数据一致性和时效性要求达到 97%、噪声、不合规异常数据占比应低于 1%。

g）遵循统一的格式标准，数据字段应清晰定义，确保字段名称、数据类型和长度等信息的准确性和一致性，定义数据加载标准方法及接口，数据加载过程应高效、快速，减少不必要的等待和延迟。

5.3 向量模型训练

向量模型训练环节包括但不限于明确向量模型的要求与规范、选择适用于向量模型训练的数据集、模型训练与优化、模型评估与调优，具体要求如下：

a）实施向量模型选择，定义模型专业领域标准要求，选择向量模型时，应需考虑模型参数规模、检索精度、向量检索性能、鲁棒性、计算资源需求等性能指标，同时重点考虑在相似应用场景下的性能表现。

b）确保所选模型可准确捕捉文本中的语义信息，同时保持较高的计算效率，以适应企业实际应用需求；

c）遵循知识数据标注标准规范，准备训练数据集，训练数据应具备多样性、代表性和规模性，确保覆盖各种可能的文本类型和语义场景；

d）选择向量模型迭代训练技术方法，逐步优化模型的参数，以提高其在文本向量表示和语义理解上的准确性，应监控模型的微调训练性能指标，包括但不限于语义相似度、向量检索性能、计算效率、鲁棒性等，应根据阶段评估指标调整训练策略，如学习率、批次大小等；

e）定义模型评估方法、指标、测试数据集、自动化工具，应对已训练的向量模型开展全面的性能评估，包括其在指标上的表现，以及与其他模型的对比，并根据评估结果，对模型开展必要的调整优化，调整优化手段包括但不限于：调整模型结构、增加正则化项等，以提高模型的泛化能力和鲁棒性，向量模型语义相似度不低于 95%，向量模型 token 输入大小不低于 512 个 tokens。

5.4 知识向量化

a）知识向量化包括但不限于知识数据块状分解、知识数据向量化、向量质量评估，具体要求如下：

b）遵循标准知识库向量数据标准规范，实施知识数据块状分解，数据块的大小应根据企业知识应用场景和需求进行设定，每个数据块应包含完整的知识单元，确保信息的连贯性和完整性，数据块之间应保持相对的语义独立性，减少信息冗余和重复；

c）以训练迭代的向量模型为基础，将以拆解的知识数据块转换为固定维度的向量表示，使用 m 均值池化或其他聚合方法得到句子或文档的向量，向量维度应适中，可充分表达知识的复杂性和多样性，尽可能保留原始知识的语义信息，确保向量能够准确表达知识的含义；

d）定义知识向量化评估规范、指标、测试数据集、自动化工具，应重点评估语义相似度、准确性、一致性、泛化能力等指标，评估向量是否准确反映了知识的核心信息。

5.5 向量知识库构建

向量知识库构建包括但不限于选择向量知识库、向量化知识库存储、向量索引构建、向量接口开放，具体要求如下：

a）实施向量知识库选择，定义向量知识库专业领域标准要求，选择向量模型时，应考虑向量知识库支持的业务需求，包括但不限于数据的规模、类型、处理速度、可拓展性及完善的安全机制；

b）实施向量化知识库存储，应保持原始数据、向量化知识数据两者均存储，采用高效的存储策略，确保向量化知识的快速存储和读取，；

c）定义高效、快速的向量索引，索引应支持动态更新，并在保障性能的前提下，应尽量优化索引的存储结构，减少占用空间建立向量索引以支持快速检索和相似度计算；

d）定义稳定性、拓展性强的知识数据向量接口，其内容应包括但不限于接口功能、调用方式、参数说明等，完善接口安全性控制，确保已授权用户才能访问和使用。

5.6 向量知识维护和更新

向量知识维护和更新包括但不限于知识更新机制、向量模型更新、错误修正与反馈，具体要求如下：

a）定义向量知识库维护与更新机制，包括但不限于建立定期更新机制，根据业务需求和数据变化频率，设定合理的更新周期，数据更新前对数据进行筛选，确保新增或修改的数据质量，避免引入错误或过时信息，对于大规模向量知识库，应采用增量更新方式，减少更新时间和资源消耗，实施版本控制策略，记录每次更新的内容、时间和操作人，以便于追踪和回滚。

b）明确模型适应性评估方法，在更新向量模型前，应对模型在新数据上的适应性进行评估，确保模型更新后的性能提升，可使用新数据进行模型再自学习与训练，在模型更新后，参照向量模型训练与评估方案，在各项评估指标实施测试，并对向量模型进行版本管理，记录模型的更新历史、性能表现及适用场景，向量知识维护周期或频率宜每周更新一次，向量模型训练宜每月更新一次；

c）建立错误识别与运营机制，应对知识库中的错误进行分类，提供知识库运行维护功能，对于发现的错误，应及时进行修正与补充，确保知识库的准确性，建立用户反馈机制，收集用户对知识库使用过程中的问题和建议，作为改进的依据，记录每次错误修正的详情，包括错误描述、修正措施、修正时间和操作人等，以便于后续审查和优化。

6 基于大模型的企业向量知识库增强检索应用

6.1 技术框架

基于大模型的企业向量知识库增强检索应用框架整合了大模型的语义理解和文本生成能力，以及向量知识库的高效检索特性，实现基于企业向量知识库的增强检索应用，主要包括：检索意图识别、知识检索与整合、内容生成与优化、安全与隐私保护 4 个应用环节，如图 2 所示。

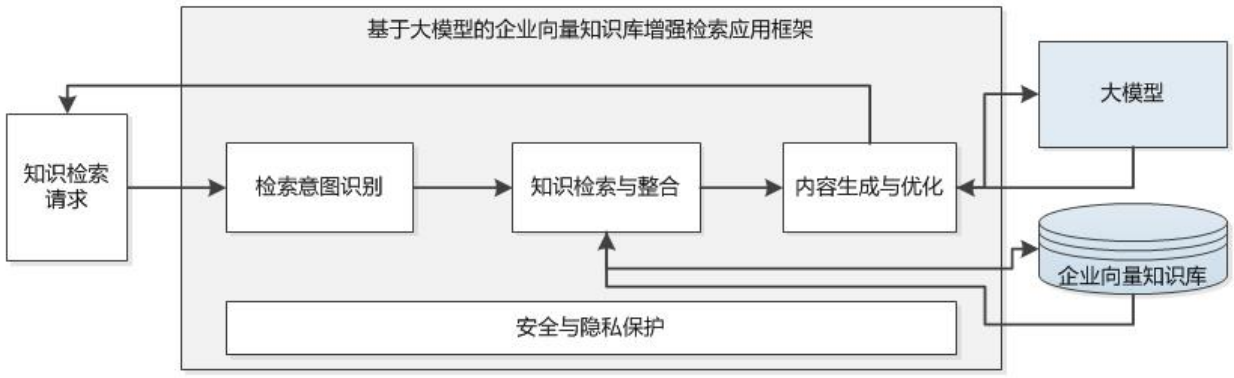

图 2　基于大模型的企业向量知识库增强检索应用框架

6.2　检索意图识别

检索意图识别模块需要准确理解用户的查询意图，为后续的企业向量知识库增强检索过程提供精准的语义支持，具体要求如下：

a）检索意图识别模型应具备高准确率，能够覆盖广泛的意图类型，且能够快速扩展到新领域；

b）检索意图识别模型需具备可解释性，能追溯到具体的训练数据和推理过程。代码需保证高可读性和可维护性，遵循标准的编码规范。标注数据的质量可追溯，并提供工具辅助标注和审核；

c）检索意图识别模型应具备快速扩展新意图类别的能力，能够快速适应新领域的需求。模型结构需支持复用和迁移，以降低新领域模型的训练成本。

6.3　知识检索与整合

知识检索与整合模块需要利用向量知识库的高效检索能力，快速找到与用户意图相关的知识片段，并将这些片段进行有效整合，为内容生成提供丰富的信息源。具体要求如下：

a）知识检索与整合模块需要具备高效的检索能力，基于向量查询快速返回相关的文档或数据。模块能够对检索结果执行多维度评估，包括相关性、权威性和时效性等，以确保所提供知识的质量；

b）知识检索与整合模块应能够对知识片段进行去重、排序和优选，优先展示最相关和最有价值的信息；

c）知识检索与整合模块应支持整合至少两种类型的异构数据源，如文本、图像、视频、语音等，以提供全面的知识视图和提升用户体验；

d）知识检索与整合模块模块还需要提供用户自定义检索策略的功能，以满足特定业务场景的需求；

e）知识检索与整合模块应具有可解释性和可追溯性，以支持检索策略、整合规则和评估模型的清晰理解和问题诊断;

f）知识检索与整合模块应提供知识审计工具，支持人工审核和优化知识库，以不断提升知识检索的准确性和相关性；

g）知识检索与整合模块的设计应具备高度的可扩展性，需要能够快速接入新的异构知识源，而无需对现有系统架构进行大规模改动。模块的检索策略和整合规则需要设计为可插拔的，以便于未来的升级和维护；

h）知识检索与整合模块应支持在线学习机制，能够根据用户互动和反馈持续优化其检索和整合模型，以适应不断变化的业务需求和技术发展。

6.4 内容生成与优化

内容生成与优化模块将检索到的知识片段通过大模型的文本生成能力转化为用户友好的、准确的答案，具体要求如下：

a）内容生成与优化模块需要利用大模型生成准确、流畅的自然语言文本，为检索结果提供清晰、有条理的摘要或答案；

b）内容生成与优化模块应能够根据用户的查询意图和检索到的知识片段，定制化生成答案内容，确保答案的针对性和个性化；

c）内容生成与优化模块应应用文本优化技术，包括语法校正、风格一致性和内容充实，以提升生成文本的质量；

d）内容生成与优化模块应支持多轮对话机制，根据用户反馈实时调整和优化生成的内容，以实现动态优化和个性化服务；

e）内容生成与优化模块应通过摘要和内容高亮等技术，提高答案的可读性和用户满意度，增强信息的突出性和易理解性。

6.5 安全与隐私保护

需要确保在检索和内容生成的过程中，用户的个人数据和企业的知识资产得到妥善保护，符合相关法律法规和企业政策，具体要求如下：

a）实施数据加密和访问控制机制，保护存储和传输中的数据安全；

b）对敏感信息进行脱敏处理，避免在检索结果中泄露；

c）提供安全审计和日志记录功能，以支持事后分析和责任追溯。

团 体 标 准

基于大模型的企业向量知识库构建及增强检索应用技术框架

T/CES 133—2024
2023 年 X 月第一版
*
北京西城区莲花池东路 102 号天莲大厦 10 层
邮政编码：100055
网址：http://ces.org.cn/html/category/17060132-1.htm
电话：010-63256990 63256997