编者按 随着 ChatGPT、DeepSeek 等生成式人工智能大模型的推出，AI 必将逐步走进百行千业万企全民。而目前大模型的幻觉问题，是影响进程的主因之一，也是基础大模型在城轨行业垂直落地应用的关键所在。构建城轨行业知识库，为各城轨企业大模型提供外挂 RAG 向量数据库等服务，是提升企业大模型训练微调效率与推理决策精准度的有效办法之一。本文提出的基于深度知识图谱，构建高质量的城轨行业知识库的相关研究，是将大模型应用到行业的基础性工作。希望行业凝心聚力，共同研究，共同建设。

doi: 10.3969/j.issn.1672-6073.2025.02.001

基于知识图谱的城轨大模型 RAG 检索增强知识库构建研究

于松伟，刘 巍，夏秀江，邵 昕，韩德志，韩晓艺
（北京城建设计发展集团股份有限公司，北京 100037）

摘 要： 当前，数据是城轨大模型落地的关键和核心养料，检索增强生成(retrieval-augmented generation，RAG)技术是城轨行业大模型建设和解决大模型幻觉问题的重要手段之一，但却因行业知识库的缺失难以充分发挥效用。本研究通过实体分类表、术语词典、属性库、实体关系表，创建分类骨架-语义基准-特征规则-逻辑关系四维架构，尤其新增实体的行业属性，突破传统知识图谱的实体 A-关系-实体 B 三元组架构，从而形成标准化与立体化的行业知识体系。基于此构建的高质量行业知识库作为 RAG 技术的核心组件，通过数据采集→结构化→向量化→知识化的链路，为大模型提供标准、可信、可溯的领域知识，显著提升城轨大模型生成内容的的可靠性和专业性，为城轨行业迈向数据驱动与知识驱动的新阶段提供核心支撑。

关键词： 城市轨道交通；人工智能；大模型；DeepSeek；RAG；知识库；知识图谱；向量数据库；数据标注
中图分类号： U231 **文献标志码：** A **文章编号：** 1672-6073(2025)02-0001-07

Constructing a Retrieval-Augmented Generation Knowledge Base for Urban Rail Transit Large Language Models: A Knowledge Graph-Based Approach

YU Songwei, LIU Wei, XIA Xiujiang, SHAO Xin, HAN Dezhi, HAN Xiaoyi
(Beijing Urban Construction Design and Development Group Co., Ltd., Beijing 100037)

Abstract: Data plays a crucial role in the successful deployment of urban rail transit large language models (LLMs). Retrieval-Augmented Generation (RAG) technology emerges as a promising approach for developing industry-specific LLMs and mitigating hallucination issues. However, the lack of comprehensive industry knowledge bases hinders its effectiveness. This study proposes a novel framework for constructing a knowledge graph-based RAG knowledge base for urban rail transit

收稿日期：2025-03-10 修回日期：2025-03-24
第一作者：于松伟，男，硕士，教授级高级工程师，长期从事城市轨道交通设计咨询与总体管理工作，yusw@bjucd.com
基金项目：国家重点研发计划课题(2024YFB3713002)
引用格式：于松伟，刘巍，夏秀江，等. 基于知识图谱的城轨大模型 RAG 检索增强知识库构建研究[J]. 都市快轨交通，2025，38(2)：1–7.
YU Songwei, LIU Wei, XIA Xiujiang, et al. Constructing a retrieval–augmented generation knowledge base for urban rail transit large language models: a knowledge graph–based approach[J]. Urban rapid rail transit, 2025, 38(2): 1–7.

LLMs. This framework consists of four key dimensions: classification skeleton, semantic benchmark, feature rules, and logical relationships. These dimensions are implemented through entity classification systems, terminology dictionaries, attribute libraries, and entity relationship tables, respectively. By incorporating industry-specific attributes for entities, this approach goes beyond the traditional subject-predicate-object triple structure of knowledge graphs, resulting in a comprehensive and multi-faceted representation of industry knowledge. This knowledge base serves as the core component of the RAG system, providing standardized, reliable, and traceable domain knowledge through a systematic pipeline of data collection, structuring, vectorization, and knowledge representation. This process significantly enhances the reliability and domain expertise of the content generated by urban rail transit LLMs, paving the way for a new era driven by both data and knowledge.
Keywords: urban rail transit; artificial intelligence; large language models; DeepSeek; retrieval-augmented generation; knowledge base; knowledge graph; vector database; data annotation

0 引言

自 2023 年 ChatGPT 发布后，以大模型技术为代表的新一轮生成式人工智能浪潮席卷而来，引发了一场千模大战。目前，DeepSeek 已在世界范围内进入百行千业万企。城市轨道交通(简称“城轨”)行业各企业也不例外，纷纷基于 DeepSeek 开展企业大模型或各种智能体研发与部署，我国城轨行业智慧化正加速迈向数据驱动与知识驱动的新阶段。

城轨行业作为大模型的应用方，也面临着算法-算力-数据三要素的结构性问题。算力与算法可由 IT 及大模型等专业厂商主导；城轨行业尽管可通过借助(开源)基础大模型搭建起本领域的垂直大模型，但由于当前行业数据质量参差不齐且覆盖范围有限，使得数据成了大模型在城轨行业应用的核心关键。NetEval 评估集系统对 26 个基础大语言模型进行评估后发现，未注入行业数据的通用大模型对行业问题的回答准确率不足 50%[1]。城轨大模型的应用既包括一般问答等通用场景，也包括决策性专用场景，其对数据的准确性与合规性要求较高。准确性不足则易在实际应用中引发故障甚至事故；模型输出若未绑定合规依据与溯源路径，还可能面临合规风险与监管处罚。

在解决上述核心问题的众多技术中，检索增强生成技术(retrieval-augmented generation，RAG)至关重要，它由领域知识库、检索器、生成器(大模型)3 大基础组件构成。其中，城轨行业知识库作为核心组件，结构化并语义向量化存储行业知识；检索器根据用户提问，精准理解用户意图并在知识库中迅速关联语义相关数据；生成器(大模型)结合检索结果和用户问题，生成符合要求的可信内容，从而提升大模型回答的准确率。

然而，当前行业内 RAG 技术的实际应用状况却差强人意，主要存在两个层面的问题：一是知识库建设层面，知识库作为 RAG 技术的运行基础，其质量优劣对 RAG 技术性能起着决定性作用，但目前城轨行业尚缺乏一个高质量的行业知识库，各企业自建知识库普遍存在知识覆盖不全、数据源杂乱、术语不统一、语义含混、更新滞后等问题；二是知识图谱支撑层面，行业知识图谱作为知识库的神经中枢，在构建行业知识库时，担负着实体齐全、概念清晰、关联准确、业务顺畅的核心职能。倘若构建的知识图谱范围不全、质量不高，那么定会导致 RAG 技术语义关联能力弱、准确性差。

基于此，本文研究了基于深度知识图谱的城轨大模型 RAG 知识库构建，以推动各城轨大模型更好地实际应用。

1 城轨行业知识库概述

1.1 城轨行业知识库简介

城轨行业知识库是指在传统结构化知识库的基础上，基于城轨专业知识构建深度知识图谱，并实现全部数据向量化的行业知识管理系统。它由行业凝心聚力，共同完成，覆盖各层级项目的全生命周期、全专业、全场景，涉及各利益相关方的知识、信息、数据，除传统知识库的各种功能外，还能为企业智能体及大模型构建提供高质量数据集、精准知识支持以及可信推理保障等基础支撑。本文提出的城轨行业知识库构建思路与技术路线，也可为城轨企业知识库的构建提供借鉴与参考，城轨行业知识库与大模型的关系如图 1 所示。

传统知识库以全文存储和关键词检索为主，仅完成了碎片化及关键词索引构建，并未进行语义向量化，语义关联性弱；而城轨行业知识库则是在传统知识库已完成碎片化的基础上，实现了知识语义化、特征向量化，可提供向量数据检索，同时知识具备跨专业、跨生命周期、跨场景等交叉关联，知识关联性强。

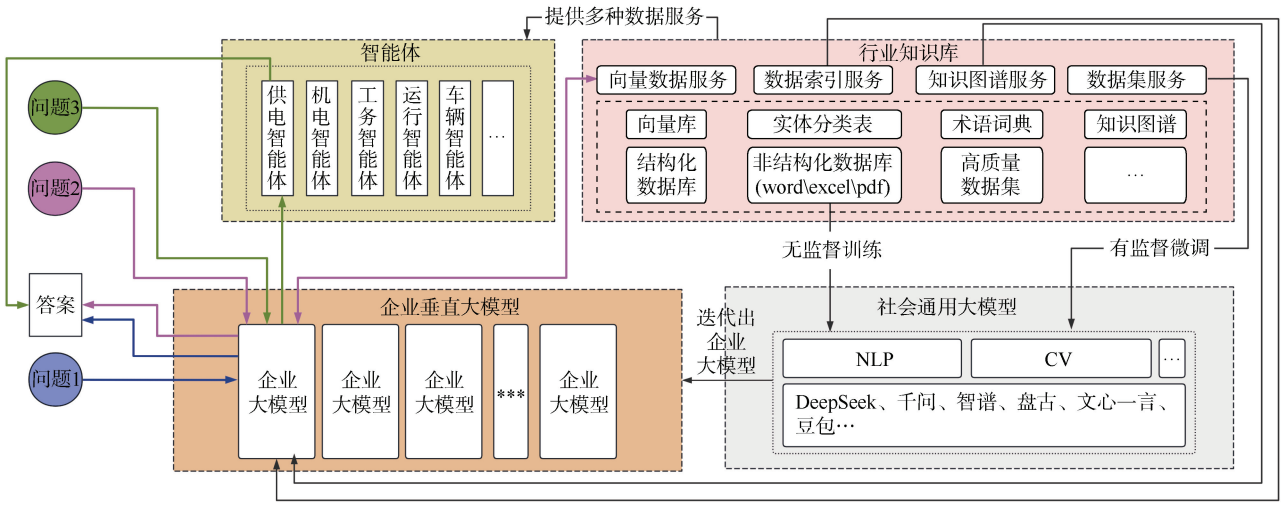

图 1　城轨行业知识库与大模型的关系

Figure 1　Relationship between urban rail transit knowledge base and large language models

城轨行业知识库可通过 RAG 技术，为城轨企业大模型提供以下服务：

1) 提升训练微调效率。Haoran Que 等指出，将法律、数学、代码、化学、音乐、医疗 6 个行业的数据，按照特定比例与社会通识数据进行融合，基于此融合数据集来训练垂直大模型，仅需投入不到社会通用大模型 10%的训练成本，即可完成训练，有效提升垂直大模型的训练微调效率[2]。

2) 提高推理准确性。腾讯优图通过构建一个涵盖 53 本医学专著和超过 38 万个医学问题的知识库，大大提升了大模型对医学问题回答的准确性[3]，法律知识库通过 RAG 技术使得 Athena 大模型在法律场景回答准确率高达 95%[4]。

3) 增强大模型安全合规性。模型输出结果自带内容溯源(如引用的标准规范条款)，可追溯每条结论的数据源头。

1.2　城轨行业知识库架构

城轨行业知识库架构，是以知识体系为内容核心，以知识库平台为技术载体，以知识单元为智能组件，构建起数据采集→结构化→向量化→知识化的完整链路，服务于城轨行业的多样化场景，如图 2 所示。

在此架构下，首先对采集到的各类数据资源进行结构化处理，再对知识体系进行向量化处理，并将其加工成数据向量库，最后在知识库平台上封装为方便调用的知识单元——智能组件。

1.2.1　知识体系

知识体系是城轨行业知识库构建的内容核心，包含知识采集与知识加工两大环节。

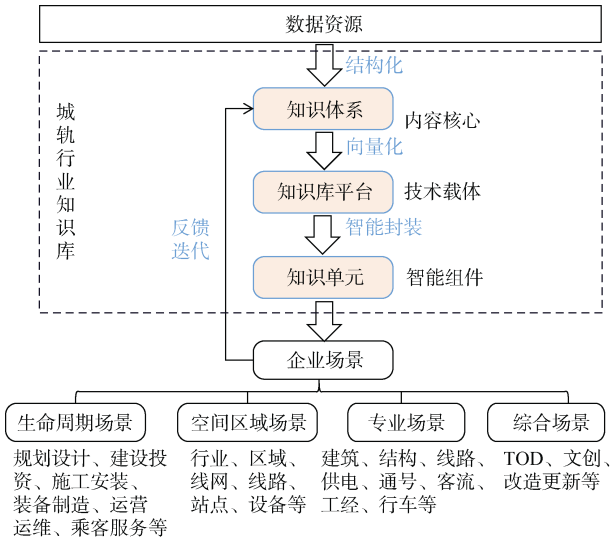

图 2　城轨行业知识库架构

Figure 2　Architecture of urban rail transit knowledge base

1) 知识采集：数据种类包括文本、图像、视频、语音等多模态数据。数据来源主要包括政府部门、行业协会、城轨相关企事业单位等行业正规渠道。

2) 知识加工：通过人机交互构建深度知识图谱，并进行数据标注，最后实现数据向量化。

1.2.2　知识库平台

知识库平台由基础模块、知识模块和应用模块 3 大模块协同构成，如图 3 所示。

1) 基础模块：提供算力支撑与硬件环境，包括智能计算集群(GPU/TPU)、分布式存储系统及机房设施，保障知识服务的高效稳定运行。

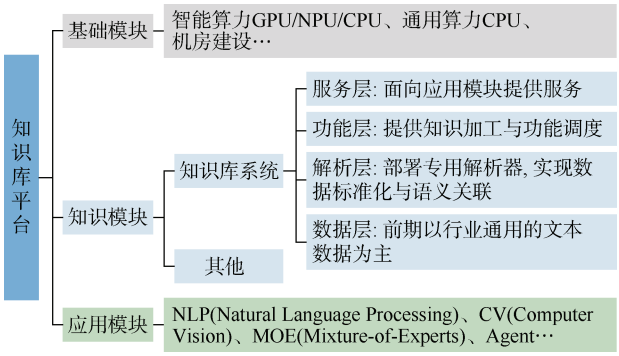

图 3　城轨行业知识库平台架构

Figure 3　Platform architecture of urban rail transit knowledge base

2) 知识模块：主要由知识库系统与其他系统构成，其中知识库系统包含数据层、解析层、功能层、服务层，实现知识的全流程管理。

3) 应用模块：面向大模型的智能服务接口，支持自然语言处理(NLP)、计算机视觉(CV)、智能体(Agent)等模型的训练微调与复杂推理。应用模块通过与这些大模型的紧密协作，将城轨行业知识库中的知识转化为实际生产力，满足行业多样化的业务需求。

1.2.3　知识单元

知识单元是基于城轨行业分类特点封装的可复用知识模块，由文档、术语词典、实体分类表、实体关系库、知识图谱、结构化数据库、向量化数据库、数据集等知识模块构成，如团标单元、年鉴单元、百科全书单元、统计分析报告单元等。城轨企业通过标准化 API 接口可直接调用知识单元，快速形成模型训练微调或推理决策所需的可靠语料与数据。知识单元支持动态更新，确保与行业政策、标准规范、实时资讯等内容的同步迭代。

1.3　城轨行业知识库的核心价值

城轨行业知识库的核心价值在于知识质量高与覆盖范围大，二者共同构成 RAG 技术高效运行的基石。

1.3.1　知识质量高

城轨行业知识库的知识质量高主要体现在数据来源的权威性和数据表征的准确性两方面。

1)数据来源的权威性指城轨行业知识库需要聚合各类经过审验的可靠数据，如标准规范、行业年鉴、协会文件、分析报告等。通过纳入这些来源明确的数据，可为知识库奠定坚实基础，同时要坚决避免未经审验的信息，以免对知识库的严谨性和可靠性造成损害。

2) 数据表征的准确性包含术语标准化和知识结构化两部分。术语标准化要求关键概念的定义严格遵从相关标准规范，从而杜绝术语存在歧义现象。鉴于城轨行业的复杂性以及协同作业的特性，一致、精准的术语定义对于各方顺畅沟通与高效协作至关重要。知识结构化指借助深度知识图谱，让各个知识点相互关联且层次清晰，这不仅便于知识的存储和管理，更能让 RAG 技术在检索知识时，迅速、精准地定位与问题相关的知识块，为大模型提供全面、准确的知识支撑。

1.3.2　覆盖范围大

城轨行业知识库需实现全生命周期、全空间层级、全利益相关者、全专业、全制式、全场景的覆盖。可按照急用先行的原则从行业共性文本知识开始，范围逐步扩展。

全生命周期确保项目各阶段的知识连贯，全空间层级满足从宏观系统到微观单元的多元知识诉求，全利益相关者助力各方协同合作从而提升城轨系统综合效益，全专业促进跨专业知识的交融与关联，全制式适配各类城轨项目独特需求，全场景保障灵活应对城轨交通复杂多变的业务场景。

2　深度知识图谱构建

2.1　定义与特点

本文的深度知识图谱是传统知识图谱在城轨行业的深化与拓展，由实体分类表、术语词典、属性库、实体关系表组成，以解决行业知识离散、语义混乱、特征未辩、逻辑缺联等问题，实现城轨行业知识的精准表达与深度应用。

相较于传统知识图谱，深度知识图谱的优势体现在以下 3 个方面：

1) 表征精准：传统知识图谱聚焦于实体 A-关系-实体 B 的三元组结构，以简单直观的平面化方式描绘知识间的基本关联，往往缺乏特定行业的术语解释、实体属性。深度知识图谱采用分类骨架→语义基准→特征规则→逻辑关系的四维架构，基于城轨行业特性、语义规范，尤其是强调了实体的行业属性，极大地增强了知识体系的标准化与立体化。

2) 广度扩展：深度知识图谱全面覆盖城轨全生命周期、全空间层级、全利益相关方、全专业、全制式、全场景，知识覆盖范围更为广泛。

3) 深度突破：深度知识图谱实现细粒度语义关联解析，从宏观系统到微观单元，实体分类层层深入，对知识进行细致剖析；同时设置动态约束，关系附加时效性等。

2.2　实体分类表

实体分类表是深度知识图谱的层级化骨架，通过

树状编码体系实现城轨实体的逐级分类与唯一标识。实体包括术语类与非术语类。术语类涵盖国家标准、行业标准、团体标准、(部分)地方标准中的专业术语；非专业术语类，指行业词汇与通用词汇。该表的构成要素包括分类框架和编码规则：

1) 分类框架：依据一心三轴四圈层[5]划分为 A(城轨一心类)、B(城轨三轴类)、C(城轨四圈层类)三大类，如表 1 所示；

表 1　城轨交通系统要素—— 一心三轴四圈层

Table 1　Elements of urban rail transit system: five subsystems as core, three axes of development, and four interactive spheres

框架	子系统	具体组成
一心	基础系统	包括隧道、桥梁、车站等站前设施
	运行系统	包括车辆、轨道、供电系统、信号系统、通信系统、综合监控系统、在线监测系统、车辆段、控制中心等站后设施
	服务系统	乘客信息系统、通风空调、照明、给排水、消防、电扶梯、屏蔽门、售检票系统、应急疏散指示、安检系统、问询、招援、导向、能源管理系统、大数据底座及平台等
	管理系统	运营调度指挥管理系统、运维管理、投融资管理、建设管理、资产管理、企业管理等
	经营系统	包括广告、通信、商业、TOD、物业开发、文旅开发、产业链投资等资源开发
三轴	时间向度轴	包括投资立项、规划设计、施工安装、装备制造、运营管理、维修改造等阶段
	空间向度轴	包括单元设备(系统)、单点车站、单线线路、城市线网、区域四网、全国行业六大层级
	利益相关方向度轴	包括乘客、政府、城轨企业、项目直接参建方、项目间接相关方、社会第三方
四圈层	功能层	包括提升出行效率、拓展城市空间、促进产业集聚、助力绿色减碳、营造品质生活、支持投资拉动等功能
	知识层	指由自然、科学、技术、工程、产业、经济、社会等相关知识所构成的集成性知识体系
	目标层	以经济、社会和环境的综合效益最大化为目标的
	交互层	始终与外围环境存在着人流、物流、能流、资金流、信息流的交互，尤其是以土地高效利用为核心的城市功能交互

2) 编码规则，采用字母搭配数字层级码，以字母标识大类，如 A、B、C 类，数字编码则用于逐级细化层级，像从 A-01 进一步细分到 A-0101……01，以此确保每个实体 ID 在全局范围内独一无二。原则上，在 ABC 三类中，实体不存在重复情况。

例如，实体接触网的编号为 A-0101010101，清晰归属于城轨一心类 A-运行系统 01-供电系统 0101-牵引供电 010101→牵引网 01010101→接触网 0101010101→架空接触网 010101010101→刚性架空接触网 0101010101 0101 这一分类，精准明确了其分类结构与上下类关系。

2.3　术语词典

术语词典是深度知识图谱的重要语义基准，通过标准化解决行业术语多义性问题，术语定义必须引用相关标准规范。

2.4　属性库

属性库是深度知识图谱的实体特征库。属性库由编号、标签、类型、行业属性、基础属性五类特征构成，形成实体特征描述的完整框架。

1) 编号(ID)：层级化编码(如 A-0101010101)是实体(接触网)唯一标识，体现分类归属；

2) 标签：实体的中英文对照名(如实体接触网的标签为接触网-catenary)，确保人机交互一致性；

3) 类型：采用实体分类表中的父类编码(如实体接触网的类型为牵引网的 ID，即 A-01010101)，以此限定继承范围；

4) 基础属性：包括词性、情感、同义词(如接触轨与第三轨)、近义词(如牵引变电所与牵引降压混合变电所)及元数据(如版本号 V2.3、状态有效)等，并动态管理。

5) 行业属性：对每一个术语或特殊名词的个体属性进行解释，以明确实体的功能。每个行业属性是一个属性组，组内属性一方面要考虑与上下父类与子类的继承关系，另一方面还要考虑全生命周期各场景中的特征表现。

示例：如实体接触网的编号(ID)为 A-0101010101，标签为接触网-catenary，类型为 A-01010101(此为其父类牵引网的 ID)；行业属性有①供电连续性：电分段？电不分段？②电压等级：直流 1 500 V？直流 750 V？等等；基础属性相关标注为词性名词，情感中性，各种元数据等。

2.5　实体关系表

实体关系表构建实体间的知识逻辑网络，包含 10 类基础关系及关系属性等。

实体关系可根据 10 类基础关系进行构建，如表 2 所示。考虑到实际的具体任务与行业特性，这些关系类型并非固定不变，可依据需求调整或扩展，不必局限于这 10 种。

比如，还可以设置方向性属性(单向关系与双向关系)，轨道交通建设(实体 A)促进城市发展(实体 B)，反过来，城市发展(实体 B)促进轨道交通建设(实体 A)，这两者就是一种双向关系；添加权重属性用来表明关系强度，像关键关系或次要关系，如新线规划(实体

A)的关键权重是线路途经人口密集区(实体 B)；还有时间有效性属性用以记录关系的时间范围，如某一标准规范的有效时间等。

表 2 城市轨道交通实体关系类型

Table 2 Urban rail transit entity relationship types

关系名称	示例说明
因果关系	高峰时段客流量激增(实体 A)导致列车发车调度调整(实体 B)
组成关系	线路网络(实体 A)由线路 1(实体 B)、线路 2(实体 C)、线路 3(实体 D)等组成
从属关系	城市轨道交通(实体 A)从属于城市公共交通(实体 B)
位置关系	换乘枢纽(实体 A)位于三条线路交叉节点(实体 B)
时间关系	2023 年城轨年鉴(实体 A)早于 2024 年城轨年鉴(实体 B)发布
相似关系	车站直梯(实体 A)与自动扶梯(实体 B)都是乘客运输设备
对立关系	列车满载率高(实体 A)与乘客舒适度达标(实体 B)存在对立
依赖关系	列车自动驾驶安全性(实体 A)依赖 ATO 系统可靠性(实体 B)
交互关系	乘客(实体 A)通过移动支付(实体 B)完成票务交互
等同关系	第三轨(实体 A)与接触轨(实体 B)同义

3 基于深度知识图谱的高质量城轨知识库构建

对城轨交通行业而言，高质量知识库构建的重点在于知识体系建设，而其难点又在于其中的数据标注。因平台部署与知识单元封装属于成熟技术，本文不再赘述。研究团队参照国家发展改革委员会的相关文件[6-7]，将数据标注解构为数据筛选、数据清洗、数据分类、数据注释、数据标记、质量检验共六大步骤，这是基于深度知识图谱实现知识向量化，并最终构建城轨行业知识库的过程。

3.1 数据筛选

根据数据类型，城轨数据可以分为文本数据、图像数据、视频数据和音频数据。

根据知识通用性，城轨知识可以分为两通两私：社会通识知识、行业通用知识、企业私域知识和个人私有知识。以城轨文本数据为例进行分层说明，如表 3 所示。

数据筛选是知识体系建设的初始阶段，其核心目标是从海量城轨数据中提取符合大模型训练与推理需求的高价值语料等原始数据，确保数据质量与业务的相关性。

3.2 数据清洗

数据清洗是根据深度知识图谱的术语词典与属性库对文本进行纠错去重、降噪、消歧等结构化处理，确保数据格式与语义的规范性，将原始文本转化为规范化的数据。

表 3 城轨文本数据分层

Table 3 Classification of text data for urban rail transit

分层	来源渠道	示例
社会通识知识	免费获取	网上公开信息、免费数据集等
	购买类	国家及各地的统计年鉴、市志等
行业通用知识	协会积累	团体标准、统计年鉴、百科全书(城轨部分)、报告、文件、装备认证、科技成果、协会网站与人民城轨公众号等
	行业公开	论文、专著、专利、软著、行业研究报告(白皮书、蓝皮书等)、行业政策、管理办法、国家标准、行业标准、媒体报道等
企业私域知识	企业	规划设计文件、勘察资料、投融资报告、施工组织、施工方案、施工管理、装备与设备手册、运营数据、管理数据、经营数据、资产数据、运维保养规程及数据、咨询评估报告、第三方监测数据、企业标准类、企业研究报告等
个人私有知识	个人	汇报 PPT、工作笔记、工作总结、会议记录、调研资料、邮件、聊天记录、培训素材等

3.3 数据分类

数据分类分为两步，第一步是基础分类，第二步是行业分类。基础分类以内容分类、词性分类等为主，行业分类是在基础分类的基础上，根据深度知识图谱进行细化。

1) 基础分类：包括内容分类(篇、章、节、段、块)，词性分类(名词、动词、形容词等)、情感分类(文本的积极、中性、消极情感倾向等)等[8-10]。基础分类可通过机器完成。

2) 行业分类：对城轨交通的实体基于实体分类框架划分层级，确定实体编号(ID)、标签、类型。行业分类以人工主导、机器为辅助，通过人机交互完成。城轨交通的实体一般属于名词或名词性描述，包括概念、命名实体，时间实体，数值实体，事件实体，评价实体等。

3.4 数据注释与标记

数据注释与标记的重点对属性标注。基础属性由机器完成，行业属性则通过深度知识图谱，通过人工注释与标注，并通过机器转化为合适维数的向量化数据。需要注意的是，向量化数据并不是维数越高越好，向量的维数需根据任务需求、数据规模和计算资源进行权衡。

3.5 质量检验

质量检验分为基础检验和行业检验，形式上为人

机协同，即先通过机器检验再人工抽样检查。基础检验是通过机器确保标注与语义保持一致性。行业检验是基于深度知识图谱检验标注是否符合实体分类表、术语词典、属性库、实体关系表，并同步检查新增标注与历史版本兼容性的对比，避免知识冲突。

4 城轨行业知识库在 RAG 技术中的核心作用

4.1 提升 RAG 技术价值发挥

知识库为 RAG 技术提供城轨专业知识支持，快速定位城轨复杂场景的关联知识，从而解决城轨大模型的知识泛化、输出欠准等幻觉问题。

4.2 驱动 RAG 语义解析升级

基于深度知识图谱的知识库，使检索器能够将口语化或非标准以及深度专业的用户提问，与知识库进行语义对齐，转换为标准、规范、专业的术语，消除自然语言中的歧义表达，使得生成器(大模型)输出精准结果。

4.3 保障 RAG 能力动态演进

知识库的动态更新使 RAG 技术能力与行业发展同步，并持续适应城轨复杂巨系统的演变。

4.4 支撑 RAG 输出溯源闭环

知识库使得 RAG 生成的结果合规、可信。知识库支撑生成器(大模型)每条输出均可关联具体的知识来源，形成输出结果-知识来源-原始文件的溯源路径。

5 结束语

城轨行业知识库是在传统知识库仅进行碎片化的基础上，结合蕴含行业属性的深度知识图谱，构建的全文向量化数据库。它包括原始文档、术语词典、实体分类表、实体关系库、知识图谱、结构化数据库、向量数据库和高质量数据集等多种数据库。作为各城轨企业一个外挂的行业知识库，除提供传统知识库关键词查询检索等功能外，还将有效提升城轨大模型的训练与微调效率，同时作为城轨大模型 RAG 技术的核心组件，将大大提高城轨大模型检索增强及其输出结果的精准度，包括为各种智能体决策提供多种数据服务。

未来课题组还将围绕知识丰富-技术适配-工具开发三大方向继续研究。

1) 知识丰富：融合图像、音频、视频等多模态数据，不断提高知识库的知识规模与范围、数量与质量。

2) 技术适配：面向 Agentic RAG 特性优化知识库架构，根据城轨行业特点和用户习惯，借助智能体技术优化 RAG 调用流程与交互逻辑，进一步提高 RAG 技术在大模型应用中的效率与价值。

3) 工具开发：在通用数据标注工具的基础上，二次开发符合城轨行业特点的城轨数据标注工具，提高行业知识库建设的效率与质量。

致谢 感谢中国城市轨道交通协会宋敏华副会长、仲建华副会长的行业指导，以及中国知网张宏伟总经理的技术指导！

参考文献

[1] MIAO Yukai, BAI Yu, CHEN Li, et al. An empirical study of NetOps capability of pre-trained large language models [EB/OL]. 2023: 2309.05557. https://arxiv.org/abs/2309.05557v3.

[2] QUE Haoran, LIU Jiaheng, ZHANG Ge, et al. D-CPT law: domain-specific continual pre-training scaling law for large language models[EB/OL]. 2024: 2406.01375. https://arxiv.org/abs/2406.01375v1.

[3] WU Ji, LIU Xien, ZHANG Xiao, et al. Master clinical medical knowledge at certificated-doctor-level with deep learning model[J]. Nature communications, 2018, 9(1): 4352.

[4] PENG Xiao, CHEN Liang. Athena: retrieval-augmented legal judgment prediction with large language models[EB/OL]. 2024: 2410.11195. https://arxiv.org/abs/2410.11195v1.

[5] 于松伟，刘巍，仲莹萤. 基于复杂巨系统理论的城市轨道交通可持续发展综合评价[J]. 都市快轨交通，2023，36(5): 1-10.
YU Songwei, LIU Wei, ZHONG Yingying. Comprehensive evaluation of sustainable development of urban rail transit based on open complex giant system theory[J]. Urban rapid rail transit, 2023, 36(5): 1-10.

[6] 国家数据局. 关于向社会公开征求《数据领域名词解释》意见的公告[EB/OL]. (2024-10-21)[2025-01-05]. https://mp.weixin.qq.com/s/uMcMcauaM6Hy0E-3vJMvyw.

[7] 国家发展改革委. 关于促进数据标注产业高质量发展的实施意见(发改数据〔2024〕1822 号)[A/OL]. (2025-01-13)[2025-02-20]. https://www.ndrc.gov.cn/xxgk/zcfb/tz/202501/t20250113_1395643.html.

[8] 蔡莉，王淑婷，刘俊晖，等. 数据标注研究综述[J]. 软件学报，2020，31(2): 302-320.
CAI Li, WANG Shuting, LIU Junhui, et al. Survey of data annotation[J]. Journal of software, 2020, 31(2): 302-320.

[9] 马鹤桐，王序文，沈柳，等. 医学知识标注体系设计与系统构建[J]. 中国卫生标准管理，2023，14(21): 1-4.
MA Hetong, WANG Xuwen, SHEN Liu, et al. Medical knowledge labeling system design and annotation system construction[J]. China health standard management, 2023, 14(21): 1-4.

[10] 王霞，徐向东，周光华，等. 健康医疗大数据标签体系构建方法研究[J]. 中国卫生信息管理杂志，2021，18(2): 189-193.
WANG Xia, XU Xiangdong, ZHOU Guanghua, et al. Research on the construction method of the tag system for health care big data[J]. Chinese journal of health informatics and management, 2021, 18(2): 189-193.

（编辑：王艳菊）