基于大模型的工业互联网物模型智能生成技术研究

Research on Intelligent Generation Technology of Industrial Internet Thing Model Based on Large Language Model

蒋　维[1]，王延红[1]，朱　亮[1]，范　斌[2]，刘　慧[1]，王秀娟[1]，王　均[1]（1. 联通数字科技有限公司，北京 100032；2. 中国联合网络通信集团有限公司，北京 100033）

Jiang Wei[1], Wang Yanhong[1], Zhu Liang[1], Fan Bin[2], Liu Hui[1], Wang Xiujuan[1], Wang Jun[1] (1. China Unicom Digital Technology Co., Ltd., Beijing 100032, China; 2. China United Network Communication Group Co., Ltd., Beijing 100033, China)

摘　要：

针对工业互联网平台中物模型构建效率低下等问题，提出了一种基于大语言模型（LLM）和检索增强生成RAG框架的物模型智能生成技术。首先，介绍了工业互联网平台发展背景及物模型构建的现状和挑战；然后，详细阐述了基于LLM + RAG的物模型智能生成技术原理；最后，以大型工业锻造设备为例，展示了该技术在格物Unilink工业互联网平台上的应用效果。研究表明，该技术能够有效提高物模型生成的效率，为工业互联网平台的应用推广提供有力支撑。

关键词：

工业互联网；物模型；大语言模型；检索增强生成

doi：10.12045/j.issn.1007-3043.2024.11.004

文章编号：1007-3043（2024）11-0019-06

中图分类号：TN915.1

文献标识码：A

开放科学（资源服务）标识码（OSID）：

Abstract：

In order to improve the efficiency of thing model constructing in Industrial Internet platform, an intelligent generation technology based on Large Language Models (LLM) and Retrieval-Augmented Generation (RAG) framework is proposed. At first, the development background of Industrial Internet platforms, the current status and challenges of thing model construction are introduced. Then the technical details of LLM+RAG-based intelligent model generation are elaborated, Finally, taking the thing model of a large scale industrial forging equipment as an example, it demonstrates the application effects of this technology on the Gewu Unilink Industrial Internet platform. The research indicates that this technology can effectively improve the efficiency of model generation, providing robust support for the development of Industrial Internet platforms.

Keywords：

Industrial internet; Thing model; Large language model; Retrieval-augmented generation

引用格式：蒋维，王延红，朱亮，等. 基于大模型的工业互联网物模型智能生成技术研究[J]. 邮电设计技术，2024(11)：19-24.

0 引言

2024年7月18日举行的中国共产党第二十届中央委员会第三次全体会议强调了健全促进实体经济与数字经济深度融合制度的重要性。会议提出了加快新型工业化的步伐，培育和壮大先进制造业集群，并推动制造业向高端化、智能化、绿色化方向发展的目标。此外，会议还强调了加快新一代信息技术的全方位、全链条普及应用，发展工业互联网，并打造具有国际竞争力的数字产业集群。

收稿日期：2024-10-16

在这一背景下，工业互联网平台作为工业互联网的中枢，被视为新型工业体系的操作系统[1]。为了支持这一平台的发展，国家出台了一系列扶持政策，旨在不断壮大包括综合型、特色型和专业型在内的工业互联网平台体系。中国联通聚焦建设网络强国、数字中国主责，拓展联网通信、算网数智主业，加大了数字基础设施的建设力度，并推出了中国联通工业互联网平台——格物Unilink平台。该平台以格物连接与设备管理平台为核心底座，成功入选国家级双跨平台，跻身我国工业互联网平台的国家队。

随着工业互联网技术的迅猛发展，联网的设备数量和数据类型呈现爆炸式增长。为了有效管理和利

用这些设备产生的海量数据，构建一个准确、高效的物模型变得尤为关键。物模型不仅定义了设备的功能和属性，还规范了设备间的通信协议和数据交换格式。面对大规模设备集成的挑战，传统的依赖人工编写的物模型构建方法显得效率低下且容易出错。为了应对这一挑战，本研究依托格物平台，提出了一种基于大型语言模型（LLM）和知识库检索增强生成（RAG）框架的工业互联网物模型智能生成技术架构。该架构旨在通过智能化的手段提高物模型生成的效率和准确性，以支持工业互联网平台的发展和物联网技术的应用。

1 背景现状

工业互联网的基础是工业物模型。工业物模型是一种数据模型，能够有效地管理设备、处理数据并提供智能命令，是实现设备智能化和互联互通的关键技术[2]。单个格物工业物模型通常包括以下几个关键组成部分。

a）属性（Property）。用于描述设备的状态或特征，例如一个传感器测量的温度值、压力值，或者一个控制器的状态。

b）服务/方法（Service/Action）。定义了设备可以执行的操作或方法，这些操作通常需要一些输入参数，并可能产生一些输出结果。例如，启动或停止一个机器。

c）事件（Event）。描述设备在特定情况下会上报的信息，这些信息可以是设备的告警、状态变化或任何需要被外部系统感知的事件。

格物工业物模型的设计兼顾了普适性、复杂性、国际化、可插拔性、安全开发、快速调试、高可靠性、可回滚性、可适配性以及统一交互协议等多个方面。多个物模型彼此互不影响，有助于解决工业场景中复杂的设备建模问题。通过物模型，可以构建数字孪生体，实现物理实体的数字化，从而促进软件与硬件的解耦，并实现智能化的物理设备生产和运维。

中国联通的格物工业物模型已在多个领域得到验证和示范应用，其中汽车制造437个、纺织印染95个、高端装备17个、消防安全57个，支撑设备接入量超71万台，解决了企业重复构建、生产设备种类多、数据要素杂等难题，助力用户智能制造升级。然而，随着物模型规模的快速扩张，如何高效构建物模型成为关键问题。工业物模型的构建过程高度依赖专业领域的设备专家，这些专家需要具备机械原理、自动化控制、计算机等多个领域的专业知识，同时，还要熟悉如数控机床标准、消防设备等相关行业标准和国家标准。

随着以ChatGPT为代表的诸多大模型出现[3]，人工智能发展出了一个新的解决问题范式，这种范式基于Next Token Prediction任务，学习并掌握了互联网上几乎所有知识，具备强大且与人类媲美的理解、认知、推理能力[4-6]。大模型在工业领域的应用正逐渐成为推动工业智能化的关键力量[7-9]。它们通过增强理解、生成和泛化能力，帮助工业企业处理海量数据，挖掘数据背后的规律和趋势。目前，大模型在工业领域的应用主要分为2个方向：一是提升场景模型的泛化能力，以提高模型的适用性；二是改变应用交互方式，如通过自然语言对话和内容生成能力，生成文档和报表等[10]。在生产制造、研发设计和经营管理3个领域，大模型的应用场景正在不断拓展。其中，在研发设计领域，大模型可以优化设计过程，提高研发效率；在生产制造领域，大模型可以拓展智能化应用的边界；在经营管理领域，大模型可以作为助手提升经营管理水平。因此，为了赋能企业建设工业互联网平台，提升项目中的物模型构建、验证、校准等过程的效率，本研究结合大模型的知识学习、知识理解和知识表达等能力，融合工业知识，进行物模型自动化构建技术研究，从而实现企业设备模型的快速构建及应用。

2 基于LLM+RAG的物模型智能生成技术架构

针对工业领域复杂的物模型快速生成需求，如果直接用大语言模型LLM来生成，通常是行不通的，主要原因有2点。

a）大模型幻觉。LLM中的幻觉现象指的是模型生成与现实世界事实不一致或毫无意义的信息。幻觉的表现形式可以分为内在幻觉和外在幻觉，其中内在幻觉指的是模型输出与输入内容相冲突，而外在幻觉指的是模型生成的信息无法通过输入内容验证。幻觉的成因与数据收集、训练和推理过程中的多种因素有关，如数据源的不一致性、模型训练中的偏差、不完美的表示学习和错误的解码策略等[11-12]。

b）模型本身缺少垂直领域知识。诸如工业、农业垂直领域的私域知识由于种种客观原因，未能上传到互联网上供各种大模型学习，因而各种大模型无法准确理解和生成垂直领域知识；此外，即使大模型学到

部分垂直领域知识，但垂直领域知识经常更新，因而大模型依旧无法准确理解最新的垂直领域知识，从而达不到用户想要的生成效果[13]。

为了减轻幻觉，研究人员提出了多种方法：在数据层面，可以通过人工标注、自动筛选和训练数据的改进来减少幻觉；在模型层面，可以通过改进模型结构、训练方式和推理策略来控制幻觉。同样地，对于模型垂直领域知识欠缺问题，研究人员也提出了通过增加垂直知识并微调等方法。但这些方法都存在成本高、迭代更新慢、模型出现遗忘等问题。

近来，随着大模型应用的不断深入，一系列通过知识外挂检索的方式来降低幻觉、补充垂直领域知识成为主流方法，该系列方法具备成本低、效率高、知识更新快等优势。主流方法有2类，分别是知识库检索增强生成（RAG）[14]和图检索增强生成（GraphRAG）[15]，二者对比来看，RAG相对更加高效且成本更低。因为GraphRAG本身依赖于知识图谱的构建，而知识图谱构建的难度、成本都很高[16]，且知识更新相对较慢。

本文提出的基于LLM+RAG的工业互联网物模型智能生成技术，首先需要收集和沉淀诸多工业垂直领域知识，然后基于LLM强大的语义理解能力、文本内容生成能力以及文本向量模型的检索能力进行物模型智能生成。

2.1 工业知识库收集与构建

如图1所示，本节主要介绍将收集到的行业物模型库、行业标准等数据，按照主题等纬度进行分类与整理，然后将整理好的数据进行有效录入与编辑，通过定期维护和物模型应用反馈对知识库内容进行更新与优化。

a）知识分类与整理。收集行业标准、已积累的行业物模型库与物模型构建标准等数据，然后对数据进行人工标注，提取关键信息和概念。根据主题、领域和用途等维度，将知识系统化地组织和分类以便大模型检索增强。

b）内容录入与编辑。将整理好的知识内容以统一的文本范式录入到知识库中，在录入之前，使用Minhash进行全局文档去重[17]。在录入的过程中，要注意保持内容的准确性和完整性，并对表格和图片等内容适当地添加文本类描述信息，以提高内容的可读性。

c）定期维护与更新。知识库是一个持续发展的过程，需要定期对知识库进行内容更新、结构优化和性能提升等操作，以确保知识库的时效性和可用性。

2.2 基于BGE模型的知识库向量检索机制构建

为了提升物模型生成的准确性与效率，本研究采用中文版BERT-based Generalized Embedding（BGE-zh）模型构建向量检索机制。BGE-zh模型作为一款先进且开源的文本嵌入模型，能够将非结构化的文本数据转换为高维空间中的稀疏向量表示，从而为相似性检索提供了一种有效的技术手段。

BGE-zh模型支持多语言，且针对中文做了较强的优化，其跨语言能力全面领先。如表1所示，BGE-zh模型在文本处理、检索精度以及资源使用情况等方面均超越了其他同类模型（包括OpenAI text embedding

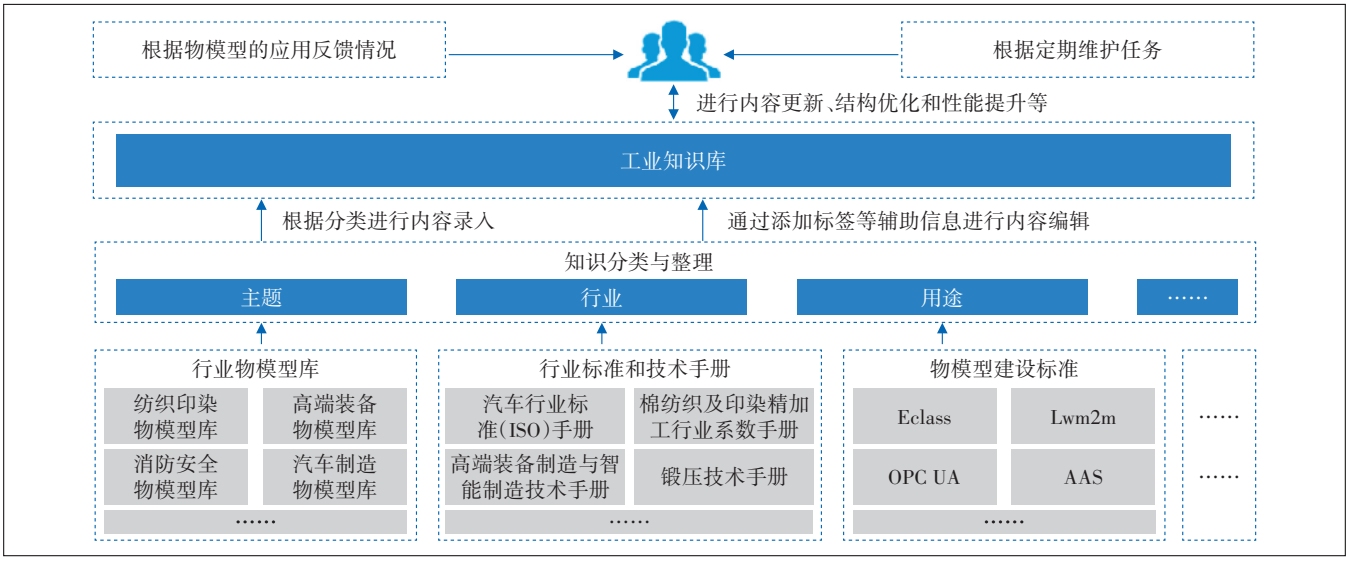

图1　工业知识库收集与构建

表1　主流文本Embedding模型性能对比

模型名称	Model Size/ Embedding Dim	检索（8 datasets）	排序（4 datasets）	句子相似度（7 datasets）	推理（2 datasets）	分类（6 datasets）	聚类（4 datasets）	平均（31 datasets）
Text2vec	base/768	38.80	49.45	41.71	63.56	65.18	37.66	47.63
Text2vec-large-chinese	Large/1024	42.11	49.16	41.98	66.39	63.42	30.02	47.36
luotuo	Base/768	44.40	54.29	39.41	63.14	65.28	44.39	49.37
OpenAI TextEmbedding 002	Large/1536	52.00	49.25	40.61	69.56	67.40	45.69	53.02
M3E-base	Base/768	56.91	59.33	48.15	63.99	70.27	47.67	57.10
M3E-large	Large/1024	54.75	59.65	48.32	64.29	71.27	48.88	57.05
BGE-zhw.o. instruct	Large/1024	70.55	64.91	50.98	76.77	72.49	50.01	63.53
BGE-zh	Large/1024	71.53	65.11	53.23	78.94	72.26	48.39	64.20

002、M3E等）。此外，BGE模型保持了同等参数量级模型中的最小向量维度，使用成本更低。因此，本文提出的物模型智能生成架构里采用BGE-zh模型进行检索增强。

由于BGE-zh模型编码的最大文本长度限制，在向量检索机制构建之前，首先要进行文本分割。文本分割主要考虑2个因素：Embedding模型的Tokens限制情况和语义完整性对整体的检索效果的影响。本研究采用固定长度分割：根据Embedding模型的Token长度限制，将文本分割为固定长度（512个Tokens）。这种分割方式会损失较多语义信息，故通过在头尾增加一定冗余提示来缓解。

其次，需要对分割后语义大致相近的文本块进行去重，与文档去重一样，本研究也采用Minhash的方式进行去重[17]。

最后，使用BGE-zh模型对去重后的文本块编码成Embeddings，并存入向量数据库，以供后续问答检索。本研究采用的是开源的Milvus向量库，该向量库专为大规模向量相似性搜索设计，支持高并发查询和插入，适合需要高性能和可扩展性的产品开发，与PG-Vector、Pinecone等其他向量数据库相比，Milvus提供了更丰富的功能和更好的扩展性。

2.3　基于LLM+RAG的物模型智能生成

如图2所示，本节以工业领域的一个“25MN快速锻造设备”为例，详细介绍基于LLM+RAG的物模型快速智能生成技术。本文采用的基座大模型是中国联通自研的元景70B大模型。具体生成算法流程如下。

Step 1：对于用户提出的“创建25MN快速锻造设备物模型”问题T_{query}，使用与2.2节相同的BGE-zh文本向量模型，将其编码成问题文本向量E_{query}。

Step 2：将E_{query}输入到向量库中进行相似度计算并检索，对检索到的候选内容集合通过公式（1）进行过滤筛选得到集合$E=\left\{E_{candidate}|\left\langle E_{query},E_{candidate}\right\rangle \geq \theta\right\}$，由于文本向量都是归一化过的向量，因此，计算向量相似度时，使用欧式距离函数等价于余弦函数。

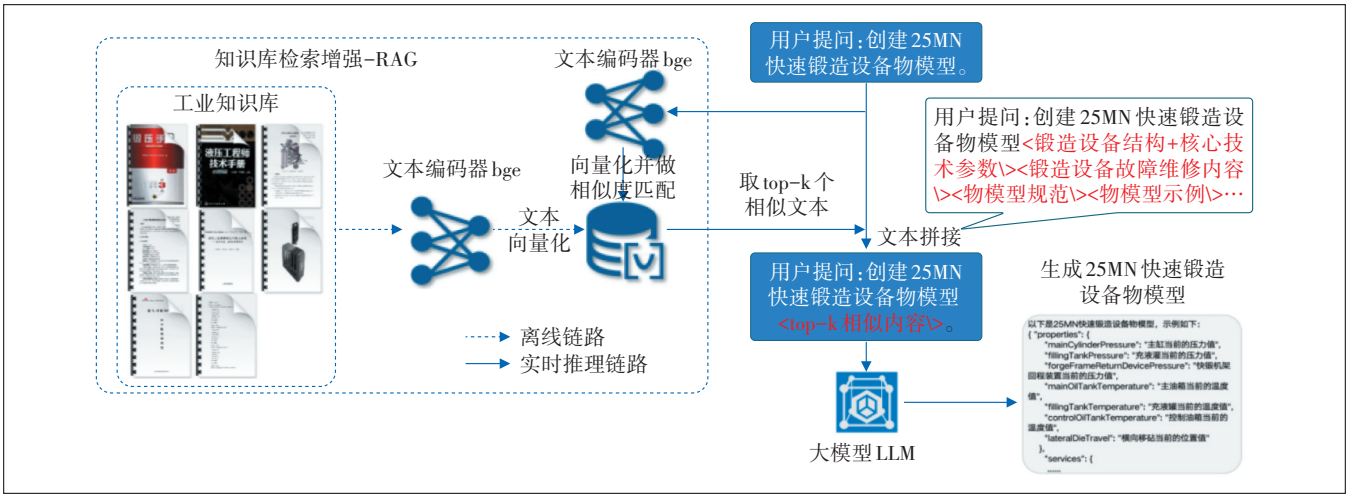

图2　基于LLM+RAG的物模型智能生成

$$\text{Similarity}\left(E_{\text{query}},E_{\text{candidate}}\right)=\left\langle E_{\text{query}},E_{\text{candidate}}\right\rangle \geqslant \theta \quad (1)$$

Step 3：对筛选后集合 E 中元素根据相似度从高到低进行排序，并取 top-k 个元素得到最终向量集合 $E_{\text{final}}=\left\{E_{\text{top1}},E_{\text{top2}},...,E_{\text{topk}}\right\}$。

Step 4：根据知识向量库中的“文本-向量”映射关系，找到 $E_{\text{final}}=\left\{E_{\text{top1}},E_{\text{top2}},...,E_{\text{topk}}\right\}$ 对应所有的文本块，并与 T_{query} 进行拼接得到最终文本块 $T_{\text{final}}=\left\{T_{\text{query}},T_{\text{top1}},T_{\text{top2}},...,T_{\text{topk}}\right\}$，达到对原始 Query 文本的增强效果，完成 RAG 算法流程，具体如图 2 所示。

Step 5：将 RAG 检索增强后的 $T_{\text{final}}=\left\{T_{\text{query}},T_{\text{top1}},T_{\text{top2}},...,T_{\text{topk}}\right\}$ 作为 Prompt 输入到基座大模型元景 70B 大模型中，并生成对应格式的物模型明细文件。

3　物模型智能生成效果

物模型生成只是工业互联网平台中设备接入管理的一个环节，物模型生成的最终应用目标是让设备以符合格物平台规范标准的方式接入进来，并自动上报数据，为后续基于上报数据进行设备全生命周期运营提供数据基础。

如图 3 所示，本节将基于格物平台，将设备物模型智能生成、物模型自动导入、创建设备、设备上报数据代码的下发部署、设备上报数据等流程串联起来，展示物模型智能生成的最终应用效果。格物平台自动导入 LLM+RAG 智能生成的物模型，导入过程中会自动检测物模型的有效性，然后根据物模型创建产品、设备以及设备对应数据上报代码。本研究中通过设备仿真的方式完成设备数据的自动上报，仍以“25MN 快速锻造设备”为例，其物模型智能生成的设备数据实时上报效果如图 4 所示。

与人工生成方式对比，单个物模型生成的时长从几个小时缩短到 1~5 min、耗时降低了近 2 个量级，考虑到知识库收集构建、向量库构建等开发工作量，整体设备接入开发效率提升近 50%，验证了基于 LLM+RAG 智能物模型生成技术架构的有效性。这种生成架构有效地提高了大模型对专业领域知识的理解、推理和快速更新能力，使其能够快速生成更加准确、专业、符合用户需求描述的物模型，从而显著提高格物平台的设备接入效率。

4　总结

本文针对工业互联网平台中物模型构建的效率和准确性低下的问题，提出了一种基于大语言模型（LLM）和知识库检索增强生成（RAG）框架的智能生成技术架构。该架构首先构建了包含行业标准、行业设

图 3　格物平台中物模型智能生成的完整应用链路

图4 基于物模型智能生成的设备数据实时上报效果

备知识和已有物模型的工业知识库,并利用BGE-zh模型实现了知识库的向量检索机制。在此基础上,结合中国联通自研的元景70B大模型,通过RAG技术实现了物模型的智能生成。以“25MN快速锻造设备”为例,详细阐述了物模型智能生成的算法流程,包括问题向量化、相似度计算、候选内容筛选、文本拼接和模型生成等步骤。最后,在格物平台上展示了从物模型生成到设备接入的完整应用流程,验证了该技术在提高物模型构建效率上的有效性。本研究为工业互联网领域的智能化发展提供了新的技术路径,具有重要的理论意义和实践价值。

参考文献:

[1] 赵春苗,王黎萤,蔡纵,等. 企业工业互联网标准化与数字创新绩效[J]. 技术经济,2024,43(8):101-113.
[2] 章刘成,张芯溢. 工业互联网赋能制造业数字化转型[J]. 商业经济,2024(5):78-81.
[3] MINAEE S,MIKOLOV T,NIKZAD N,et al. Large language models: a survey[EB/OL]. [2024-01-20]. https://arxiv.org/abs/2402.06196.
[4] 赵京鹤,童辉,卫芳芳. 数据资源体系在大数据中的应用[J]. 中国自动识别技术,2021(1):53-56.
[5] 陶晓英. 基于混合架构的大语言模型智能问答系统研究[J]. 邮电设计技术,2024(5):48-55.
[6] AN Z Y,DING X Z,FU Y C,et al. Golden-retriever: high-fidelity agentic retrieval augmented generation for industrial knowledge base [EB/OL]. [2024-01-20]. https://arxiv.org/abs/2408.00798.
[7] 康保斌,王刚,杨雷,等. 面向产业链上下游企业的基于LLM的垂直领域智能客服适配方法研究[J]. 制造业自动化,2024,46(6):7-12,100.
[8] 芦存博,左璇,金博,等. 大模型在工业安全领域的应用研究与探索[J]. 新型工业化,2024,14(7):85-95.
[9] 陈亚盛,蒋礼蔚,单敏,等. 审计大模型的构建及应用研究[J]. 审计研究,2024(4):139-149.
[10] 崔爽. AI大模型成企业智能化转型重要推手[N]. 科技日报,2023-09-11(6).
[11] FARQUHAR S,KOSSEN J,KUHN L,et al. Detecting hallucinations in large language models using semantic entropy[J]. Nature,2024,630(8017):625-630.
[12] WU S Y,XIONG Y,CUI Y F,et al. Retrieval-augmented generation for natural language processing: a survey[EB/OL]. [2024-01-08]. https://arxiv.org/abs/2407.13193.
[13] 张凯,涂志莹,陆展,等. 面向大规模定制协同生产的工业机理建模[J]. 郑州大学学报(理学版),2023,55(2):1-9.
[14] 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 System. Red Hook:Curran Associates Inc.,2020:9459-9474.
[15] EDGE D,TRINH H,CHENG N,et al. From local to global: a graph rag approach to query-focused summarization[EB/OL]. [2024-01-08]. https://arxiv.org/abs/2404.16130.
[16] 黄思蓓,张磊. 工业互联网安全知识图谱设计研究[J]. 自动化仪表,2021,42(12):90-92,99.
[17] DUBEY A,JAUHRI A,PANDEY A,et al. The llama 3 herd of models[EB/OL]. [2024-01-08]. https://arxiv.org/abs/2407.21783.

作者简介:
蒋维,高级工程师,博士,长期从事人工智能和工业互联网平台研发工作;王延红,高级工程师,学士,长期从事工业互联网平台规划工作;朱亮,工程师,博士,长期从事人工智能研发工作;范斌,工程师,学士,长期从事工业互联网平台技术、行业应用研究工作;刘慧,硕士,长期从事物联网平台研发工作;王秀娟,学士,长期从事物联网平台规划工作;王均,学士,长期从事工业互联网平台设计工作。