



工程与应用

## 基于通信网络大模型的智能运维 人机交互系统研究与应用

韩赛, 范凤霞, 马家福, 王泽林, 徐博华, 王光全, 唐雄燕  
(中国联合网络通信有限公司研究院, 北京 100044)

**摘要:** 随着网络规模的扩大和5G应用的爆发式增长, 网络管理运营面临新的需求和挑战。运营效率直接影响网络使用效能和服务质量, 亟须通过智能化手段提升网络管理运营水平, 减少传统低效、重复性工作。随着AI技术和通信网络的深度融合, 引入通信网络大模型成为提升网络管理运营效率的关键路径。对此, 研发了一套基于大语言模型的智能运维人机交互系统。该系统通过大小模型协同机制, 集成了知识问答、人机交互、数据分析与方案生成等功能。该系统在现网的部署应用表明, 其不仅能够显著提升运维效率, 降低运维成本, 还能通过预测性维护减少网络故障的发生, 提升用户体验, 增强企业竞争力。该系统具备高度可复制性与推广适应性, 具有广泛的应用前景与实用价值。

**关键词:** 大语言模型; 智能运维; 大小模型协同; 智能问答; 通信网络

**中图分类号:** TP393; TN915.07

**文献标志码:** A

**doi:** 10.11959/j.issn.1000-0801.2026020

## Research and application of intelligent operation and maintenance human-computer interaction system based on large language models for communication networks

Han Sai, Fan Fengxia, Ma Jiafu, Wang Zelin, Xu Bohua, Wang Guangquan, Tang Xiongyan  
Research Institute of China United Network Communications Co., Ltd., Beijing 100044, China

**Abstract:** With the continuous expansion of network scale and the explosive growth of 5G applications, new demands and challenges are encountered in network management and operations. Since operational efficiency directly influences network utilization and service quality, it becomes imperative to enhance the level of network management through intelligent means and reduce traditional inefficient and repetitive tasks. As AI technology becomes deeply integrated with communication networks, the introduction of large-scale communication network models is recognized as a key pathway to promote network management and operations. In response, an intelligent operational human-computer interaction system based on large language models (LLMs) was developed. This system integrated capabilities such as knowledge-based question answering, human-computer interaction, data analysis, and solution generation through a collaboration mechanism between large and small models. Deployment and application of the system in live



networks demonstrated that it not only significantly improved operational efficiency and reduced maintenance costs, but also minimized network failures through predictive maintenance, thereby enhancing user experience and strengthening corporate competitiveness. The system was designed with high replicability and adaptability, indicating broad application prospects and practical value.

**Key words:** large language model, intelligent operation and maintenance, collaboration between large language and lightweight models, intelligent Q&A, communication network

## 0 引言

随着数字化转型的深入推进,各行业纷纷借助智能化技术提升生产效率,降低人力成本。通信运营商亦顺应这一潮流,逐步由传统网络服务提供商向云网服务及数字服务提供商转型,其业务范围已由通信技术(commun-ication technology, CT)延伸至信息与通信技术(information communication technology, ICT)<sup>[1-2]</sup>。随着网络规模的不断扩大及5G应用的爆发式增长,网络管理与运维面临前所未有的挑战。传统的被动运维模式通常由客户投诉触发,依赖人工操作,往往难以实现跨专业协同,导致故障定位效率低、网络资源利用率不足。在全面数字化、自动化和智能化的趋势下,亟须构建从数据采集、分析、故障诊断、修复到优化的一体化闭环运维体系,从而将运维人员的角色从“流程执行者”转变为“流程监督者”,全面迈向主动运维新阶段<sup>[3-4]</sup>。

随着人工智能(artificial intelligence, AI)技术的持续进步,其与通信网络的融合已成为推动网络管理模式革新和运维体系升级的关键力量。AI技术具有强大的数据挖掘、模式识别和自动推理能力,近年来在网络运维领域得到了广泛的研究与实践应用<sup>[5]</sup>。具体来看,AI在通信网络中的应用主要体现在以下3个方面。

(1) 网络故障检测与预测:通过对历史告警数据、性能指标和用户行为的学习, AI模型能够准确识别潜在故障模式,实现对网络异常状况的早期发现和自动预警,从而推动运维模式从传统的“事后响应”转向“事前预防”,显著缩短故

障修复时间<sup>[6-7]</sup>。

(2) 资源优化与调度:在无线通信、数据中心及边缘计算等场景中, AI被广泛应用于频谱资源分配、功率控制、负载均衡等任务。与传统基于人工规则的优化算法相比, AI模型能够根据实时数据动态调整配置,从而提高网络资源利用率与系统性能<sup>[8-9]</sup>。

(3) 用户行为建模与体验保障:通过用户画像分析、流量预测与体验质量(quality of experience, QoE)评估等方法, AI可协助运营商识别关键业务、保障高价值用户的体验,从而实现更加精细化的网络管理与服务分层<sup>[10-11]</sup>。

尽管AI技术在通信网络中取得了初步成果,但仍面临一些挑战:一是跨专业场景中的知识孤岛问题,传统机器学习模型在专业之间难以实现知识迁移<sup>[12]</sup>。二是多源异构数据融合能力不足,系统在面对复杂故障链条时缺乏全局建模与统一推理能力。三是人机交互体验不佳,现有系统更多依赖菜单式操作,缺乏对自然语言的深度理解与响应。在此背景下,大语言模型(large language model, LLM)的预训练-微调范式和强大的语义理解与生成能力,为通信网络的智能运维提供了新的思路。LLM具备跨模态、跨场景的知识泛化能力,能统一处理来自网络的设备日志、运维文档、结构化指标等多源数据,并实现对运维过程的全流程建模<sup>[13-14]</sup>。借助对自然语言的理解和生成能力, LLM实现了对用户意图的精准识别和多轮对话式交互,为一线工程师提供“类专家系统”式的问答支持,帮助新手快速掌握技能,辅助专家处理疑难问题。LLM可以与小模型、知

识图谱等模块协同工作,实现“模型调用模型”“指令驱动分析”等复杂任务。借助函数调用(function call)机制,LLM不仅能生成答案,还能触发网络状态查询、故障定位、报表生成等实际操作,极大增强了模型的执行力与实用性<sup>[15]</sup>。

现有运维系统在一定程度上满足了用户的基本需求,但在交互体验和智能化方面仍存在明显不足。未来的运维系统应更加注重用户体验,为用户提供更为直观、友好的交互界面,并支持多样化的交互方式,以满足用户在不同场景下的需求。同时,系统还应具备更高的智能化水平,能够自动分析网络状态,预测潜在问题,并提供相应的解决方案,从而减轻运维人员的工作负担,提高运维工作效率和质量。

随着AI技术与通信网络的深度融合,引入通信网络大模型实现智能运维,不仅能够显著提升运维效率、降低人力成本,还能借助预测性维护减少网络故障,从而改善用户体验,增强企业竞争力。面对通信网络日益增长的复杂性和用户需求的多样性,基于通信网络大模型的智能运维已成为应对挑战、把握机遇的关键手段,更是推动通信行业智能化转型、提升服务质量和用户体验的必由之路<sup>[16]</sup>。

综上所述,LLM的引入有望突破传统AI在通信运维中“碎片化、工具化”的应用瓶颈,构建以理解为核心、以推理为手段、以交互为纽带的下一代智能运维系统。在复杂异构网络日益普及、运维成本持续攀升的今天,基于LLM的智能运维系统被认为是推动通信行业向更高水平智能化发展的核心抓手,具备广阔的研究前景和现实应用价值。

本文主要围绕通信网络大语言模型在智能运维中的研究与应用展开。

## 1 与大模型相关的技术方案

LLM凭借其强大的自然语言理解、推理和生

成能力,正成为新一代智能运维系统的核心技术。为实现通信网络领域的知识问答、人机交互、方案生成、数据分析等复杂功能,需要结合LLM的通用能力与通信行业的专业特点,构建完整的模型支撑技术体系<sup>[17]</sup>。为此,本文提出了一种面向通信网络运维的大模型技术架构,包括模型基础架构、知识增强机制、大小模型协同调度、多模态交互、任务驱动与系统集成等关键部分。

### 1.1 大语言模型的基础架构与训练策略

当前主流LLM多基于Transformer架构。该架构由Google于2017年提出,具备优秀的特征提取与上下文建模能力,已广泛应用于文本生成、问答、摘要、代码生成等任务<sup>[18]</sup>。LLM通常采用“预训练+微调(pretraining+fine-tuning)”的方式,即先在大规模通用语料上进行语言建模预训练,再在特定领域数据上进行微调,从而具备迁移泛化与行业适配能力。具体而言,首先在大规模通用中文语料(如维基百科、新闻、对话等)上进行语言建模预训练,使模型掌握基本的语义理解能力;随后在通信领域的高质量行业语料(如设备手册、标准文档、运维操作规范等)上进行指令微调(instruction tuning),显著提升模型在专业问答、指令生成等任务上的性能。针对特定任务需求,可进一步引入监督微调(supervised fine-tuning, SFT)、基于人类反馈的强化学习(reinforcement learning from human feedback, RLHF),以及参数高效微调方法(如P-Tuning/LoRA),使模型既能保持语言生成的一致性,又能快速适配特定任务指令<sup>[19]</sup>。

本文采用参数规模在70亿级别以上的中文大语言模型(如ChatGLM、DeepSeek等)作为底座模型,并结合通信领域专业语料进行进一步微调,以增强模型在通信网络与运维场景下的语言理解与推理能力,使其具备“懂网络、懂运维”的专业特性。



## 1.2 提示工程与任务引导机制

为提升系统对多源异构数据的感知与处理能力，大模型集成了多模态输入机制，支持文本、语音、结构化指标、拓扑图等多种输入形式。通过对这些数据的统一嵌入表示与跨模态联合建模，大模型能够实现语义层面的对齐与融合，从而满足通信运维中“文本描述+图表指标+系统命令”的复杂输入场景需求。

在实际交互过程中，智能运维系统还引入了基于提示工程（prompt engineering）的任务驱动机制<sup>[20]</sup>。提示工程在此不仅是模型输入的格式化手段，更是连接“运维意图识别”与“任务执行逻辑”的桥梁。通过为不同任务设计结构化、高鲁棒性的提示模板（prompt template），系统能够结合当前用户意图与历史上下文，动态构建高质量的模型输入。该机制可有效引导大模型理解复杂问题、规划处理路径，并实现与小模型的自动联动或基于检索增强生成（retrieval-augmented generation, RAG）的知识增强调用，从而提升模型输出的相关性与准确性。为了让模型更加稳定且准确地响应用户问题，本文设计了一套完整的提示工程机制。

### （1）任务分类式提示模板设计

根据通信运维的任务类型，将提示模板划分为如下4类，每类对应一种典型的运维场景。

其一，基础知识问答类（如“什么是OTN?”）提示模板示例：

你是通信网络专家，请基于专业知识回答用户的问题：“{问题内容}”。

其二，指标分析类（如“查询某端口利用率”）提示模板示例：

请分析设备“{设备名称}”下端口“{端口号}”在最近“{时间范围}”的利用率情况。

其三，故障诊断类（如“某基站无服务原因”）提示模板示例：

基于以下告警信息、网络拓扑与历史操作记

录，请判断该基站无服务的可能原因，并给出处理建议：{告警数据片段}。

其四，预测预警类（如“某核心网元是否需要扩容”）提示模板示例：

请结合核心网网元用户历史数量，预测“{网元名称}”未来90天内是否存在容量瓶颈风险。

### （2）提示模板与自动联动协同

针对结构化任务，如端口利用率查询、拓扑自动识别等，提示工程不仅要激活模型的语言理解与规划能力，还需要明确地触发特定小模型的调用。例如，在执行“IPRAN（IP radio access network）与基站自动拓扑关联”任务时，提示不仅要描述问题，还要嵌入调用函数的语义信号，如请判断“{IPRAN}”与“{基站}”间的端口连接关系，并绘制拓扑结构图。调用方法：`get_topology(ipran_id, bts_id)`。

结合自动联动机制，智能运维系统能够自动生成参数并调度后端模块执行任务，返回结果后再由大模型进行结构化归纳与语言优化输出。

### （3）多轮对话与上下文记忆提示机制

在智能运维场景中，运维工程师往往会通过多轮交互逐步定位问题。因此，提示工程需要具备对话上下文缓存与问题状态追踪能力，以实现“任务链式思维”引导。

### （4）示例驱动少样本提示机制

为提升模型在专业任务中的泛化能力与输出稳定性，系统构建了基于通信领域高质量样例的少样本提示（few-shot prompting）模板库，为模型提供参考格式与回答风格。少样本提示示例如下。

示例输入：

问题：什么是IPRAN？

示例输出：IPRAN（IP radio access network）是一种基于IP技术的无线回传解决方案，主要用于连接基站与核心网……

示例输入：

任务：查询某端口利用率并生成报表。

示例输出：该端口在过去7天的平均利用率为75%，高于阈值70%，建议检查上下游链路瓶颈……

通过在提示中引入“样例引导”，大模型能更好地适配任务语言风格与信息结构，并输出更具可读性与专业性的内容。

### 1.3 通信知识库与语义向量库构建

为支撑高质量的问答与推理能力，智能运维系统构建了覆盖运营商通信网络中无线、IP地址、传输、核心网等专业网络的通信知识库。该知识库采用3层架构进行组织。

(1) 文档库：存储文本格式的知识文档，包括原始的PDF、DOC、HTML等文件。

(2) 结构化知识库：存储从文档中提取的实体、关系和属性等信息，涵盖设备、指令、参数、告警等关键实体及其关联关系。

(3) 语义向量库：存储知识片段的向量化表示，采用BERT、SimCSE等嵌入模型对文本进行向量化，以实现高效的语义检索。

在问答过程中，系统先将用户提出的问题转化为向量，再基于语义相似性从知识库中检索最相关的知识片段，并将其与问题一同输入大模型，实现基于RAG的问答机制。该机制能够有效减少大模型生成错误信息（“幻觉”现象），提升回答的知识准确性与可解释性。

为保证检索的准确性与效率，语义向量库在构建过程中遵循以下标准，并经历以下优化过程。

#### (1) Embedding模型选择标准

Embedding模型作为语义向量库的核心，其选择直接影响了知识检索的覆盖度与精度。Embedding模型主要从以下4个维度进行选择。

- 覆盖度：优先选用在大规模中文通用语料和技术语料预训练中表现优异的模型（如SimCSE、ERNIE、BERT等），确保模型具备广泛的语义理解基础。
- 领域适配性：在通信运维领域的专有语料

（如设备手册、标准规范、运维日志、告警记录等）上进行微调，增强其对行业术语、缩写及操作语义的建模能力。

- 效率与可扩展性：结合大规模知识库的实时检索需求，选择在推理效率与参数规模上具有平衡性的模型，确保在GPU/CPU混合部署环境中仍能实现低时延响应。
- 鲁棒性与稳定性：通过对比实验，验证模型在短文本（如告警信息）、长文本（如标准文档）以及含噪声输入（如拼写错误或格式不规范）等多种场景下的检索性能，确保其在复杂场景下仍能保持较高的准确率。

#### (2) 相似性算法优化过程

为提升语义检索的效率与准确性，对相似性度量方法进行了多层次优化，具体过程如下。

- 基础方法：采用余弦相似度作为主要度量方式，其在衡量高维稠密向量的语义相似性方面表现稳定。
- 索引优化：引入高效的近似最近邻（approximate nearest neighbor, ANN）搜索框架，如Faiss、HNSW，实现对百万级别向量库的毫秒级检索，并支持GPU加速以进一步降低查询延迟。
- 混合度量机制：在特定场景（如跨专业拓扑匹配）下，将余弦相似度与欧氏距离进行加权融合，以提升检索鲁棒性和边界区分能力。
- 动态权重调整：针对不同类别的知识片段（如标准规范、设备手册、告警信息等），引入类别权重因子，使关键性知识在检索结果中具有更高优先级。
- 上下文增强检索：在多轮对话或复杂任务场景中，通过查询扩展（query expansion）与历史上下文关联，提升检索的召回率和语义一致性。



## 1.4 大小模型协同机制

在通信运维任务中，很多操作具有“输入结构化—查询—输出结构化”的特点，适合由轻量模型/程序自动执行。因此，可以引入大小模型协同机制，其中，大模型用于理解与指令生成，小模型用于执行与反馈。具体而言，大模型负责意图识别、上下文理解、任务规划、结果生成；小模型是原子能力，主要负责特定任务的高效执行，如告警查询、端口利用率分析、拓扑自动识别等。

为实现模型间的协同，系统引入了函数调用机制。该机制允许大模型根据自然语言意图调用已注册的小模型应用程序编程接口（application program interface, API），如RESTful接口，从而完成结构化数据处理、数据库操作、图谱检索等任务，并将执行结果返回大模型进行内容汇总与最终展示。具体调用过程如下。

(1) 大模型首先解析用户意图，生成结构化调用请求，包括函数名称与参数表。

(2) 该请求经由系统中注册的统一API网关转发给对应的小模型。

(3) 小模型接收参数后执行具体任务（如数据库查询、端口利用率分析、拓扑图绘制等），并返回执行结果。

(4) 大模型对小模型返回的结果进行归纳总结与语言优化，生成用户可读的响应内容。

该机制不仅规范了模型间的调用流程，也确保了参数传递的一致性和结果处理的标准化。

具体的提示模板示例如下：

你是一名通信网络智能运维助手，请根据用户输入生成标准化的函数调用请求。

用户意图 {用户的问题或需求}

输出要求

1. 必须以JSON格式输出，包含两个字段

- "name": 要调用的函数名称

- "arguments": 一个参数字典，包含执行任务所需要的参数及其值

2. 函数名称必须准确反映任务类型，例如

- get\_port\_utilization（端口利用率查询）

- get\_topology\_relation（拓扑关联分析）

- get\_capacity\_forecast（容量预测）

- get\_alarm\_diagnosis（告警诊断）

3. arguments内的参数必须结构化填写，例如

```
{
  "device_name": "xxx",
  "port_id": "xxx",
  "time_range": "xxx"
}
```

4. 如果用户意图不涉及具体调用，请输出

```
{
  "name": "none",
  "arguments": {}
}
```

用法示例

用户输入：查询基站A过去7天GE0/0/1端口的利用率

大模型输出：

```
{
  "name": "get_port_utilization",
  "arguments": {
    "device_name": "基站A",
    "port_id": "GE0/0/1",
    "time_range": "last_7_days"
  }
}
```

## 2 系统方案

### 2.1 总体思想

基于网络数据和专家知识，并借助大模型，智能运维人机交互系统能够将传统运维操作流程与指令转化为直观的问答交互模式，从而减轻运营管理工作负担，提升运营管理效率，助力运营商网络向高等级的自智网络演进。基于多模态大

模型，构建了覆盖单专业与跨专业领域的通信网络知识库，训练生成面向通信网络的领域大模型，并围绕云网运营生产和管理需求，实现了多个典型场景的落地应用。该网络大模型以“懂网络、懂运营”为核心目标，是面向网络自智运营而构建的通信领域行业大模型。其具备知识问答、人机交互、数据分析和方案生成等能力，并能协同场景化小模型与知识图谱等技术，通过对话式运营运维模式，大幅提升作业效率。该模型旨在解决运维领域的复杂问题，或者降低问题的复杂度。其目标是使机器掌握通信知识，理解网络结构、运营场景与工程师意图，从而在网络规划、建设、维护、优化和运营等环节提供智能支持，进一步提升整体作业效率。

此外，系统支持语音交互、文件交互等多种方式，并将跨专业网络拓扑自动关联、端口利用率自动查询、网元容量预警等小模型原子能力与网络大模型融合，构建了一个更为便捷、高效的智能运维人机交互系统。基于通信网络大模型的智能运维人机交互系统界面如图1所示。

## 2.2 技术方案

为实现网络基础知识的智能问答，本系统通过输入无线、IP和传输等专业网络的技术文档、通用知识等，构建了一个可动态维护的知识库。

用户通过知识库管理界面对知识内容进行增加、删除、查询等操作。当用户上传新文档后，系统将自动对文档进行格式转换与语义切片处理，并利用选定的Embedding模型将文档文本转化为向量，存储至语义知识库中，以支持后续的语义检索与问答。

知识库的数据来源主要包括以下3个方面。其一，企业内部数据，包括设备文档、操作手册、历史故障案例等。其二，人工标注数据，由领域专家对知识库中的数据进行标注、校正和补充。其三，互联网公开数据，如学术论文、行业标准、行业报告、专业博客等。

在问答流程中，系统首先使用Embedding模型将用户问题编码为向量，然后通过相似性算法在知识库中检索，找出与用户问题最贴合的知识片段。检索结果中得分最高的知识库内容将与原始问题一同输入大模型，由大模型进行综合理解与语言生成，最终输出贴合用户需求的回答。基础知识问答实现流程如图2所示。

为验证大模型在网络运维中的实际效能，分别选取跨专业网络拓扑自动关联、端口利用率自动查询、网络负荷健康状况监测与预警3个现网高价值场景作为典型应用。下面将分别介绍各场景下小模型原子能力及其与大模型的协



图1 基于通信网络大模型的智能运维人机交互系统界面

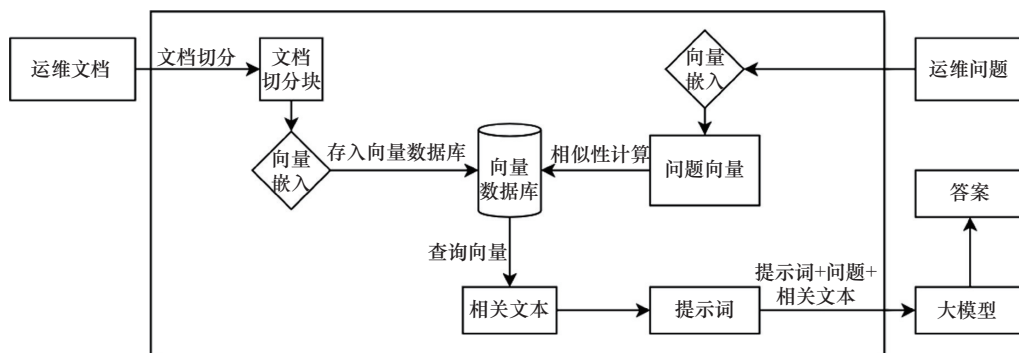


图2 基础知识问答实现流程

同机制。

实现跨专业故障自动诊断的前提是，能够自动形成跨专业网络拓扑，标识出专业内、不同专业间网络设备的逻辑连接与物理连接，进而结合多专业告警信息，准确定位网络中的故障点。资源、连接数据是准确绘制跨专业网络拓扑的基础。目前，跨专业网络资源之间普遍缺乏自动关联能力。端到端拓扑关联包括专业网络内拓扑关联和跨专业网络拓扑关联两部分。其中，专业网络内拓扑关联手段较为丰富，可通过采集专业控制器拓扑信息或通过专业网络特有协议获取，因此，本文主要针对跨专业网络拓扑关联展开阐述。针对 IPRAN 和基站、IPRAN 和光传送网（optical transport network, OTN）等不同专业组合间的异构特性，分别提出了不同的拓扑自动关联方案。

### 2.2.1 小模型原子能力—IPRAN和基站跨专业网络拓扑自动关联

针对 IPRAN 和基站设备都支持链路层发现协议（link layer discovery protocol, LLDP）的情况，本方案可以直接使用 LLDP 采集云存储网关（cloud storage gateway, CSG）设备上的 LLDP 信息，并与基站侧信息进行直接对比，实现设备间的自动关联。这种方法简便有效，但是需要设备具备 LLDP 支持能力。目前，已有大部分 5G 基站支持 LLDP<sup>[21]</sup>。

对于不支持 LLDP 的存量 3G/4G 基站，可通过

监测地址解析协议（address resolution protocol, ARP）信令的方式实现关联。具体流程为：首先，从无线网管系统获取基站相关信息；然后，与 IPRAN 侧通过 ARP 协议采集的 MAC 地址进行匹配，从而实现 IPRAN 和基站设备的跨专业自动关联。

### 2.2.2 小模型原子能力—IPRAN和OTN跨专业网络拓扑自动关联

针对 OTN 设备没有 IP 地址，并且大部分设备不支持 LLDP 的情况，本方案结合 AI 技术，采用基于端口流量曲线相似性的 IPRAN 和 OTN 互联端口匹配方法，实现跨专业互联链路的自动匹配。

以广东省某地市 IPRAN 接入 OTN 环路为例，每个环路包括 4 个相连端口：接入 IPRAN—接入 OTN—汇聚 OTN—汇聚 IPRAN。系统获取指定时间段各端口的接收和发送流速数据。在已知  $N$  个环路、 $4N$  个端口的  $4N$  个接收和  $4N$  个发送流速数据的情况下，基于相连端口的流量曲线具有相似的特点，采用时间序列相似性算法，依次从所有发送流速数据中选出和每个接收流速数据最匹配的 2 个发送数据。同时，结合 AI 算法，系统可以判断出接入 IPRAN 发送流速、接入 OTN 接收流速、汇聚 OTN 发送流速、汇聚 IPRAN 接收流速之间的相似关系，进而绘制出各个环路中 4 个端口的连接关系。各个环路中 4 个端口的相连关系示例如图 3 所示。本方案适用于单环路及多环路场景下的拓扑关联。

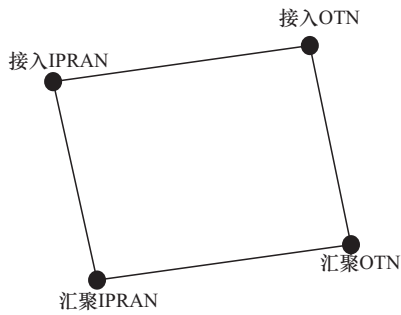


图3 各个环路中4个端口的相连关系示例

### 2.2.3 大小模型协同—跨专业拓扑自动关联

在问答场景下，系统首先调用大模型对用户输入的问题进行意图识别，并根据识别结果调用不同的方法。然后，通过相似性算法在知识库中检索，找出回答问题所需要的数据。最后，由大模型将相关数据与原始问题进行整理后输出，在前端以列表、图文等形式展示，支持下载功能。大小模型协同—跨专业网络拓扑关联实现流程图如图4所示。

本模块采用函数调用技术，该技术可以增强模型推理能力，并支持外部操作，包括信息检索、数据库访问、知识图谱搜索与推理、系统操作等工具调用场景。

### 2.2.4 大小模型协同—端口利用率自动查询

基于光功率等监测数据，系统可以自动计算各网元端口利用率，并结合大模型实现报表查询与智能问答。本模块在大小模型协同流程逻辑上与跨专业网络拓扑自动关联类似，具体流程可参考图4。

### 2.2.5 大小模型协同—网络负荷健康状况监测与预警（核心网）

除了提升运维效率，未来网络管理运营还需要具备主动预测风险的能力。云核心网作为通信

网络的关键组成部分，涉及云资源、数通等专业联动，面临扩容周期长、难以快速响应市场需求等挑战。因此，亟须通过智能预测用户规模和增长趋势，提升网络资源利用率和扩容建设响应速度。

基于用户终端数、登网数等多维指标，结合节假日用户迁移等规律，引入AI算法对5GC（包括接入和移动性管理功能、会话管理功能、统一数据管理）和IMS（如呼叫会话控制功能）等网元容量进行多元预测，实现基于短期（1~10天，1~30天）、中期（1~90天）和长期（1~180天，1~365天）用户规模和趋势预测，并能对超出门限阈值的网元容量情况及时发出预警，有效提升了网络资源利用率与扩容建设响应效率<sup>[22]</sup>。

## 3 现网实施

### 3.1 系统架构

本系统在实际部署中，围绕“智能化支撑能力、高可用运行架构、灵活可扩展机制”三大目标，构建了面向现网运维场景的整体技术架构。系统采用微服务化部署、模块解耦协同、异构环境融合等设计理念，在满足通信行业高并发、高可靠需求的前提下，实现了从数据采集、语义理解到任务执行、可视化呈现的全流程闭环支撑。

系统整体分为人机交互前端、应用服务中台、智能引擎支撑后端、运维数据中心4个运行单元，部署在山东联通骨干云平台上，具备良好的可移植性和可复制性。

### 3.2 现网实施效果

基于通信网络大模型的智能运维人机交互系统已在某省级运营商云平台完成部署。智能运维人机交互问答效果如图5所示。该系统基于网络

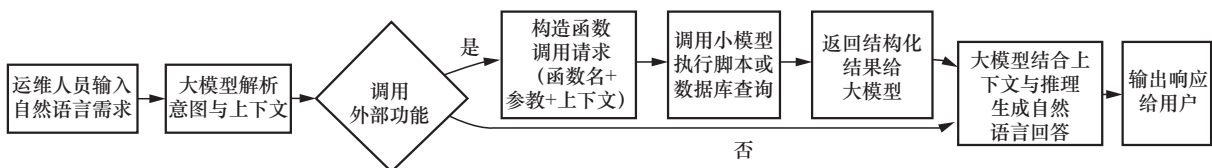


图4 大小模型协同—跨专业网络拓扑关联实现流程



请问SDH网络的常见网元有哪些

SDH(同步数字体系)网络广泛应用于电信领域的一种传输网络,其常见网元主要包括以下几类:1.“SDH传输设备”:“SDH传输设备”:如SDH传输模块、SDH传输板卡等,用于实现SDH信号的传输。2.“终端复用器(Terminal Multiplexer, TM)””:用于低速信号复用成高速信号,或将高速信号解复用成低速信号。3.“分插复用器(Add/Drop Multiplexer, ADM)””:具有插入和提取信号的功能,可以在网络中插入或提取特定速率的信号。4.“同步复用器(Synchronous Multiplexer, SM)””:用于将多个低速信号同步复用成高速信号。5.“再生中继器(Regenerator)””:用于延长SDH信号的传输距离,通过再生信号来补偿信号在传输过程中的衰减。6.“线路系统设备”:如光传输系统、同步传输系统等,用于传输SDH信号。7.“交叉连接设备(Cross-Connect Device, XCD)””:用于实现网络中不同信号之间的交叉连接。8.“保护设备”:如线路保护设备、复用保护设备等,用于提高网络的可靠性。9.“网管设备”:用于对SDH网络进行监控、配置、故障处理等操作,这些网元共同构成了SDH网络,实现了高速、大容量的数据传输。

5G承载网的业务发展趋势?

5G承载网作为5G网络的重要组成部分,其业务发展趋势主要体现在以下几个方面:1.“高带宽需求”:随着5G应用范围的拓展,如高清视频、虚拟现实(VR)、增强现实(AR)、远程医疗等,对承载网带宽需求大幅提升。2.“低时延”:5G承载网需要支持更低的时延,以满足实时性要求高的应用,如自动驾驶、工业自动化等。3.“网络切片”:5G承载网支持网络切片技术,为不同业务提供定制化的网络服务,实现差异化服务。4.“智能化与自动化”:通过引入人工智能(AI)和机器学习(ML)技术,实现网络管理的智能化和自动化,提高网络运维效率。5.“边缘计算”:5G承载网支持边缘计算,将数据处理和计算能力下沉到网络边缘,降低时延,提高数据处理的实时性。6.“网络切片与边缘计算的融合”:网络切片与边缘计算的结合,将更好地满足不同业务场景的需求,实现网络资源的灵活配置和高效利用。7.“网络切片的标准化”:随着网络切片技术的不断发展,相关标准化工作也在逐步推进,以促进网络切片技术的广泛应用。8.“安全性与可靠性”:5G承载网需要具备更高的安全性和可靠性,以保障关键业务和重要数据的安全。9.“绿色环保”:在5G承载网的建设和运营过程中,将更加注重节能减排,推动

(a) 基础知识查询

您好, 欢迎使用大模型验证, 当前场景为IPRAN和基站拓扑关联, 请输入您要咨询的问题:

滨州堡集IPRAN下有哪些基站

城市	IPRAN网元名称	IPRAN网元ID	基站网元名称
滨州	SDH_设备_82_190_A01	006-15102	82BC0030-ER-F1H01-通信支路-FL三(通信支路)
滨州	SDH_设备_82_190_A01	006-15104	82BC0030-ER-F1H01-通信支路-FL三(通信支路)
滨州	SDH_设备_82_190_A01	006-15108	82BC0030-ER-F1H01-通信支路-FL三(通信支路)
滨州	SDH_设备_82_190_A01	006-12002	82BC0030-ER-F1H01-通信支路-FL三(通信支路)

请提供拓扑图

(b) IPRAN和基站自动拓扑关联

您好, 欢迎使用大模型验证, 当前场景为IPRAN和OTN拓扑关联, 请输入您要咨询的问题:

请问接入侧IPRAN侧廊坊苏村980C-0/6/0和OTN的关系, 并生成拓扑图

接入侧IPRAN廊坊苏村980C-0/6/0的关联设备为接入侧廊坊东站OSN9600-1槽-GZ20-6口、汇聚侧廊坊OSN9600-1槽-GZ20-6口、汇聚侧廊坊CX600-X8A-4/0/0

(c) IPRAN和OTN自动拓扑关联

您好, 欢迎使用大模型验证, 当前场景为端口利用率, 请输入您要咨询的问题:

请提供一下泰安2024年8月的端口利用率数据

端口名称	资源已占用	资源利用率	利用率
1111-泰安-泰安-泰安	2	4	0.5
1111-泰安-泰安-泰安	4	8	0.5
2010-泰安-泰安-泰安	2	8	0.25
2012-中心-泰安-泰安	2	8	0.25
2017-中心-泰安-泰安	1	4	0.25
3000-泰安-泰安-泰安	1	4	0.25
3011-泰安-泰安-泰安	1	4	0.25
01-泰安-泰安-泰安	1	4	0.25
C3-泰安-泰安-泰安	1	4	0.25

(d) 端口利用率报表生成

您好, 欢迎使用大模型验证, 当前场景为网元容量预警, 请输入您要咨询的问题:

请预测2024年3月的山东SMF的容量

山东SMF在2024年3月的容量是1348.7244万用户数。

请预测上海UDM在2024年3月的容量

上海UDM在2024年3月的容量是325.153万用户数。

(e) 网元容量预警

请提供华为无线网“S1接口故障告警”的告警描述:

华为无线网专业, 告警“S1接口故障告警”的告警描述为: STAP (S1 Application Protocol) 连接在底层SCTP链路资源可用时, eNodeB向MME发起连接建立请求; 如果MME对连接请求做合法性检查不通过, 将无法建立连接; 如果MME检查通过, 当eNodeB收到MME的响应后, 发现MME(MME Code)冲突时也将无法建立连接, 当STAP协议层因数据配置错误、MME异常无法建立连接或S1链路故障时, 产生此告警。当底层SCTP链路故障, 且该状态持续16秒未恢复时, 也会产生该告警; 当底层SCTP链路资源变为可用, 且该状态持续2秒一直可用时, 则上报告警恢复。

该告警的修复建议是什么?

修复建议如下:

1. 检查S1接口配置;
2. 检查网络区域配置信息;
3. 检查基站配置;
4. 检查基站运营商配置。

(f) 告警信息管理

图5 智能运维人机交互问答效果

数据和专家知识,将运维流程和操作指令转化为自然语言问答模式,减轻了运营管理工作负担,大幅提升了运维人员的工作效率。系统具有高度可复制性,可便捷推广至集团公司及其他省级分公司,具有良好的应用推广价值。

(1) 基础知识查询:支持无线、IP、传输等网络基础知识的快速查询,为运维人员提供专业知识支持,如图5(a)所示。

(2) IPRAN与基站拓扑关联查询:支持IPRAN与基站的端口相连查询,并呈现关联拓扑,如图5(b)所示。

(3) IPRAN与OTN拓扑关联查询:支持IPRAN与OTN的端口相连查询,并呈现关联拓扑,如图5(c)所示。

(4) 端口利用率报表查询:支持端口利用率的查询、预测和报表生成,有利于优化资源配置,提高利用率,如图5(d)所示。

(5) 网元容量预警:支持网元容量的查询和预测,显著提升扩容建设响应效率,如图5(e)所示。

(6) 告警信息管理:支持多厂商的告警信息查询,并能够根据告警信息提供修复建议,提升故障响应和处理效率,如图5(f)所示。

## 4 结束语

基于大语言模型的提示工程、知识增强、多模态融合、大小模型协同等关键技术,本文提出了一套面向通信网络智能运维的技术方案,并研发了基于通信网络大模型的智能运维人机交互系统。该系统已在运营商现网环境中部署运行,能够基于网络数据和专家知识,将运维流程和操作指令转化为自然语言问答模式,减轻运营管理负担,大幅提升了运维人员的工作效率。该系统具备良好的可扩展性与可复制性,能够部署于集团公司及其省级平台,具有良好的应用推广价值。在现网应用中,该系统展现出多方面的运维效益:在故障处理方面,系统

通过智能化的告警信息管理和拓扑自动关联功能,帮助运维人员快速响应并处理网络故障,降低了故障对业务的影响;在资源配置方面,借助实时的端口利用率报表查询和网元容量预警功能,运营商能够优化资源配置,避免网络拥堵和性能下降,提高网络资源的利用率;在运维成本方面,通过引入智能化运维手段,降低了运维人员的工作量与整体运维成本,为运营商带来了显著的经济效益。

## 参考文献:

- [1] 王泽林,韩赛,张洁,等.中国联通自智网络研究与实践[J].通信世界,2022(21):42-45.  
Wang Z L, Han S, Zhang J, et al. Research and practice of China Unicom's intelligent network[J]. Communications World, 2022(21): 42-45.
- [2] Han S, Ma H B, Chen D, et al. Streaming video optimization in mobile communications[C]//Proceedings of the 2018 IEEE/CIC International Conference on Communications in China (ICCC). Piscataway: IEEE Press, 2019: 495-499.
- [3] 裴培,王爽,刘一平,等.数字化转型时代下运营商IT架构进阶之路[J].信息通信技术,2021,15(6):66-71.  
Pei P, Wang S, Liu Y P, et al. The advanced path of telecom operators' IT architecture in the era of digital transformation[J]. Information and Communications Technologies, 2021, 15(6): 66-71.
- [4] Han S, Ma H B, Zhang P, et al. Improved MPEG-4 high-efficiency AAC with variable-length soft-decision decoding of the quantized spectral coefficients[J]. China Communications, 2019, 16(10): 65-82.
- [5] 周晶,王德政,洪科.5G网络智能运维AI应用研究[J].邮电设计技术,2021(11):83-87.  
Zhou J, Wang D Z, Hong K. Research on AI application in 5G network intelligent operation and maintenance[J]. Designing Techniques of Posts and Telecommunications, 2021(11): 83-87.
- [6] Menaria V K, Jain S C, Raju N, et al. NLFFT: a novel fault tolerance model using artificial intelligence to improve performance in wireless sensor networks[J]. IEEE Access, 2020, 8: 149231-149254.
- [7] Mohammed S. Artificial intelligence in computer networks: delay estimation, fault detection, and network automation[D]. Ottawa: University of Ottawa, 2021.
- [8] Cao Y, Wang R, Chen M, et al. AI agent in software-defined network: agent-based network service prediction and wireless



- resource scheduling optimization[J]. IEEE Internet of Things Journal, 2020, 7(7): 5816-5826.
- [9] Aron R, Abraham A. Resource scheduling methods for cloud computing environment: the role of meta-heuristics and artificial intelligence[J]. Engineering Applications of Artificial Intelligence, 2022, 116: 105345.
- [10] Ribeiro H, Barbosa B, Moreira A C, et al. Customer experience, loyalty, and churn in bundled telecommunications services[J]. Sage Open, 2024, 14(2): 21582440241245191.
- [11] Banjanin M K, Stojčić M, Danilović D, et al. Classification and prediction of sustainable quality of experience of telecommunication service users using machine learning models[J]. Sustainability, 2022, 14(24): 17053.
- [12] 韩赛, 张冬月, 王泽林, 等. 跨专业承载网络智能运维研究与应用[J]. 电信科学, 2022, 38(11): 113-122.  
Han S, Zhang D Y, Wang Z L, et al. Research and applications on intelligent operations of cross-professional carrying network[J]. Telecommunications Science, 2022, 38(11): 113-122.
- [13] Zhou H, Hu C M, Yuan Y, et al. Large language model (LLM) for telecommunications: a comprehensive survey on principles, key techniques, and opportunities[J]. IEEE Communications Surveys & Tutorials, 2025, 27(3): 1955-2005.
- [14] Ma J F, Han S, Wang G Q, et al. An LLM-based cross-domain fault localization in carrier networks[C]//Proceedings of the 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC). Piscataway: IEEE Press, 2025: 731-736.
- [15] Abdelaziz I, Basu K, Agarwal M, et al. Granite-function calling model: introducing function calling abilities *via* multi-task learning of granular tasks[J]. arXiv preprint, 2024, arXiv: 2407.00121.
- [16] 韩赛, 范凤霞, 叶晓斌, 等. 面向优质用户体验的自智网络研究与应用[J]. 信息通信技术, 2023, 17(3): 34-41.  
Han S, Fan F X, Ye X B, et al. Research and application of autonomous networks for high quality user experience[J]. Information and Communications Technologies, 2023, 17(3): 34-41.
- [17] Bariah L, Zhao Q Y, Zou H, et al. Large generative AI models for telecom: the next big thing? [J]. arXiv preprint, 2023, arXiv: 2306.10249.
- [18] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing System. New York: Curran Associates, 2017: 6000-6010.
- [19] Kan K B, Mun H, Cao G H, et al. Mobile-LLaMA: instruction fine-tuning open-source LLM for network analysis in 5G networks[J]. IEEE Network, 2024, 38(5): 76-83.
- [20] Marvin G, Hellen N, Jjing D, et al. Prompt engineering in large language models[M]//Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2023. Singapore: Springer,

2024: 387-402.

- [21] Han S, Wang Z, Wang G, et al. Automatic association of cross-domain network topology[C]//IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Piscataway: IEEE Press, 2022: 1173-1178.
- [22] Han S, Li A, Zhang D Y, et al. Early warning of core network capacity in space-terrestrial integrated networks[J]. Journal of Systems Engineering and Electronics, 2024, 35(4): 855-864.

#### [作者简介]



韩赛 (1988-), 女, 博士, 中国联合网络通信有限公司研究院高级工程师, 主要研究方向为网络智能、自智网络等。



范凤霞 (1986-), 女, 中国联合网络通信有限公司研究院工程师, 主要研究方向为网络智能、自智网络等。



马家福 (1999-), 男, 中国联合网络通信有限公司研究院助理工程师, 主要研究方向为网络智能、网络大模型等。

王泽林 (1983-), 男, 中国联合网络通信有限公司研究院高级工程师, 主要研究方向为IP、云网、白盒、SDN技术等。

徐博华 (1989-), 男, 中国联合网络通信有限公司研究院高级工程师, 主要研究方向为数据中心网络、宽带城域网、下一代互联网等。

王光全 (1968-), 男, 博士, 中国联合网络通信有限公司研究院正高级工程师, 主要研究方向为通信网络的规划、新技术演进、标准制定等。

唐雄燕 (1967-), 男, 博士, 中国联合网络通信有限公司研究院副院长、首席科学家、正高级工程师, 主要研究方向为宽带通信、光纤传输、互联网、物联网、新一代网络等。