

一种 AI Agent 驱动的意图网络设计与实现

周洪利¹
ZHOU Hongli

摘要

面对网络环境日益异构化与配置需求复杂化的挑战,传统基于脚本或模板的自动化配置方法存在灵活性差、语义理解能力不足的问题。意图网络作为一种变革性的网络架构范式,通过将上层业务意图自动编译为底层网络策略,重构了网络管理模式。针对传统意图网络语义理解匮乏与大模型事实幻觉的问题,文章提出一种由 AI Agent 驱动的神经符号协同意图网络架构,该架构通过引入基于思维链的意图解析机制和语义-物理双重锚定机制从源头阻断幻觉,并构建基于 MCP 网关的内生安全与遥测闭环体系,有效解决了运维开环难题。实验结果表明,在非确定性生成任务中,意图解析精度与幻觉抑制效能上取得了显著提升。同时,端到端原型验证了底层确定性控制逻辑的闭环连通性,为智能网络管理提供了工程实践参考。

关键词

AI Agent; 意图网络; 软件定义网络; 神经符号协同

doi: 10.3969/j.issn.1672-9528.2026.05.020

0 引言

近年来,物联网、云计算等互联网业务的快速发展,促使网络流量呈现快速增长。这不仅对带宽、延迟、服务质量等指标提出了更高要求,也对网络配置方法提出了新的挑战。在传统网络架构中,由于软硬件耦合紧密、配置方式相对简单,导致灵活性不足。软件定义网络^[1]是一种新的网络架构,具有转控分离、集中控制、开放可编程等特点,提升了网络管理的灵活性。但是软件定义网络也面临着策略制定高度依

赖专家经验和自动化程度不足等问题。意图网络(intent-based networking, IBN)是一种新型的网络范式,又称为基于意图的网络或意图驱动网络,其基本思想是解析用户输入的业务需求,并将这些业务需求转化为可执行的网络配置。目前,IBN 已逐渐成为网络领域的研究热点。不过,当前多数 IBN 方案在处理自然语言意图时,仍依赖预定义的静态模板或者简单的关键词匹配,难以应对模糊、复杂的真实业务场景。

大语言模型的产生和发展,特别是 AI Agent 的出现,为意图网络的研究注入了新的动力。本文在梳理意图网络、软件定义网络、大语言模型及 AI Agent 研究现状的基础上,提

1. 江苏省未来网络创新研究院 江苏南京 211100

- [7] TUFEK N, YALCIN M, ALTINTAS M, et al. Human action recognition using deep learning methods on limited sensory data[J]. IEEE sensors journal, 2019, 20(6): 3101-3112.
- [8] KIM S, JEONG J, SEO S G, et al. Remote recognition of moving behaviors for captive harbor seals using a smart-patch system via bluetooth communication[J]. Micromachines, 2021, 12(3): 267.
- [9] LIU J, LIU H B, CHEN Y Y, et al. Wireless sensing for human activity: a survey[J]. IEEE communications surveys and tutorials, 2019, 22(3): 1629-1645.
- [10] WANG Z J, JIANG K K, HOU Y S, et al. A survey on human behavior recognition using channel state information[J]. IEEE access, 2019, 7: 155986-156024.

- [11] 甘星宇. 基于 Wi-Fi 子载波优化选择的人体动作识别研究[D]. 长沙: 湖南工商大学, 2022.
- [12] 苏健, 郑毓煌, 陈思光. CGAC: 一种基于 CSI 的人体动作识别方法[J]. 南京邮电大学学报(自然科学版), 2024, 44(6): 12-24.
- [13] DIAZ G, SOBRON I, EIZMENDI I, et al. Channel phase processing in wireless networks for human activity recognition[J]. Internet of things, 2023, 24: 100960.

【作者简介】

甘星宇(1997—), 女, 湖南汨罗人, 硕士, 研究方向: 无线感知与智能处理, email: 307725172@qq.com。

(收稿日期: 2026-03-24 修回日期: 2026-04-30)

出一种由 AI Agent 驱动的神经符号协同意图网络架构。其基本思想是大模型负责意图的解析，符号系统负责策略的校验与物理实体锚定。也就是该架构引入基于思维链的意图解析机制和语义 - 物理双重锚定机制，将解析出的业务意图与网络实体及约束进行精准绑定。实验结果显示，该架构在意图解析精度和幻觉抑制效果上都有较大的提升，这为意图网络提供了具有工程参考价值的实践方案。

1 意图网络

意图是一种高层次的声明式表述，对应的是网络预期的最终状态，它不关心具体的底层技术实现和配置步骤。意图网络便是基于这一概念构建的自动化管理范式。目前，意图网络的标准尚未完全统一。但在互联网研究任务组（internet research task force, IRTF）等权威机构的推动下，已经有一些工作取得进展。其中，IRTF 下属的网络管理研究组（network management research group, NMRG）发挥了关键作用，发布了 RFC 7575^[2]，其中明确了意图的概念，系统讨论了网络自动化场景下的策略执行机制与意图管理逻辑；之后，发布了 RFC 9315^[3]，这进一步界定了 IBN，区分了意图与服务模型、策略之间的差异，提出了 IBN 的运行原则、功能模块及生命周期，并补充了意图分类与安全考量。2025 年，NMRG 发布了 *Intent-Based Network Management Automation*^[4]，重点面向 5G 网络场景，阐述了基于意图的网络管理自动化的未来发展趋势。该草案提出了一个融合闭环控制、意图转译与管理审计的赋能框架，核心目标是将高层次业务目标转化为网络可执行策略，以支撑网络切片、IoT 数据聚合及车联网服务质量保障等需求，推动网络配置的智能化与自动化。

Li 等^[5]提出了一个通用的 IBN 体系结构，这个结构包含应用层、意图层和网络层等，如图 1 所示。

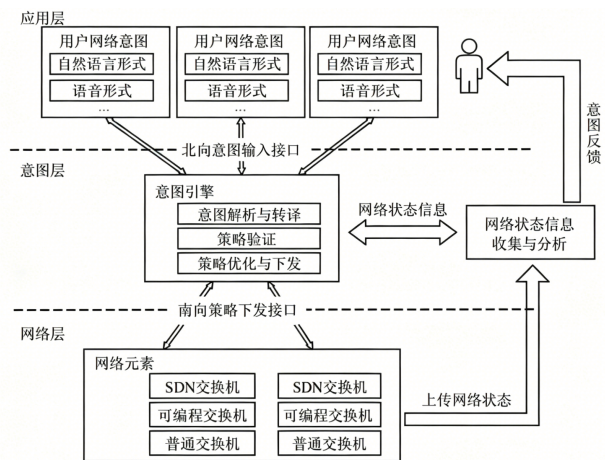


图 1 意图网络通用体系结构模型

(1) 应用层：支持用户通过自然语言、语音等形式输

入网络意图，让用户能够以声明式的方式表达业务需求。

(2) 意图层：主要是意图引擎，包含意图解析与转译、策略验证、策略优化与下发等环节，具体的工作流程是将业务意图映射为确定的网络策略，完成冲突检测与一致性验证，然后生成具体的配置指令。

(3) 网络层：主要由 SDN 交换机、可编程交换机及普通交换机等设备组成，用来执行具体的数据包转发任务。

(4) 闭环链路：用于维持网络状态与用户意图的持续一致性。网络层将实时采集的流量负载、丢包率等信息上传到收集与分析模块，然后由其对原始数据进行加工处理，之后再将分析结果反馈给意图引擎，并由意图引擎进行一致性校验。

总体上，意图网络借助 SDN 的集中控制能力以及网络分析提供的实时数据，为 AI 技术在网络中的部署创造了条件。下面将分别论述 SDN、大语言模型和智能体。

2 软件定义网络

软件定义网络的早期研究主要关注控制平面的可编程性，出现了 OpenDaylight、ONOS 等网络操作系统，并借助 OpenFlow 协议实现设备管控。但 OpenFlow 严重依赖功能固定的包处理硬件，数据平面缺乏灵活性，难以适应业务需求的快速变化。为突破这一限制，数据平面可编程技术逐步发展起来。P4 语言因其协议无关性，允许开发者自定义报文解析与匹配流程，释放了数据平面的处理潜力。目前软件定义网络已进入全可编程阶段，推动了云网络基础设施的演进。例如，Sailfish^[6]采用软硬件协同架构实现高性能云网关；POSEIDON^[7]面向超大规模弹性场景，构建了具有极致性能的虚拟网络控制器；LuoShen^[8]则基于超融合架构，解决了边缘云环境下的高并发与大单流性能瓶颈。

上述系统虽然提升了网络的传输性能与灵活性，但底层架构日益复杂，导致网络配置与管理的认知门槛明显提高。现有可编程网络提供了丰富的控制接口，然而面对海量配置参数和动态业务意图，传统的基于规则脚本的自动化方式已难以胜任，且容易引入人为错误。因此，需要引入具有更强认知能力的技术手段。这也促使研究重点从高性能转发逐渐转向认知智能化，成为将大语言模型和智能体应用于网络管理领域的主要动因。

3 大语言模型

大语言模型的发展大致经历了结构创新、规模扩张和深度推理能力的演进等阶段。Transformer^[9]是现有大模型的基础架构。在这个架构的基础上，预训练范式分为两条路径，分别是以 BERT^[10]为代表的双向编码路径和以 GPT 系列^[11]为代表的自回归生成路径。当大模型的参数规模超

过千亿后，其展现出了较好的涌现能力。LLaMA^[12]验证了开源基础模型在不同规模下的缩放定律。InstructGPT^[13]引入了基于人类反馈的强化学习，使得模型的输出可以与人类价值观对齐。

接下来，Mamba^[14]利用选择性状态空间模型，突破了Transformer的平方复杂度限制，实现了线性增长的超长上下文处理能力。Chain-of-Thought^[15]发掘了模型的分步推理潜力。OpenAI o1^[16]则引入思维链强化学习，标志着模型从快速直觉响应向深度慢思考的转变。DeepSeek-V3^[17]通过多专家模型的极致优化和多标记预测技术，提升了预训练的效率和性价比。DeepSeek-R1^[18]验证了在仅依靠纯强化学习并不依赖大规模监督微调数据的情况下，通过群体相对策略优化就能激励模型产生自我修正和长链推理能力。

4 AI Agent

AI Agent可以定义为以LLM为核心控制器、具备环境感知与自主决策能力的智能体系统，包含控制器、记忆模块、规划模块、工具调用模块等组件。其中，控制器利用LLM进行逻辑推理与决策；记忆模块维护历史状态与知识检索；规划模块负责任务拆解；行动模块输出具体指令；工具调用模块连接外部环境，将自然语言意图转化为可执行的操作。

具体到智能体的任务规划，它经历了从动态在线调整到长程闭环学习、再到推理内化的演进过程。ReAct^[19]结合“推理与行动”，解决了静态推理的局限。Reflexion^[20]引入自我反思的语言反馈闭环，使得智能体具备了试错学习的能力。Quiet-STaR^[21]则将推理能力隐式化、通用化，使得模型在没有外部反馈和特定任务的情况下，通过内生思考以增强对语言本质的理解与预测。

工具调用能力让大语言模型能够突破自身权重的限制，通过外部媒介获取实时知识并执行复杂指令。从Toolformer^[22]利用自监督学习建立API调用范式开始，研究者后续通过Gorilla^[23]解决大规模API适配与实时文档同步的挑战，并利用ToolLLM^[24]中的深度优先搜索决策树提升智能体在多工具协同任务中的推理深度。面对工具生态日益碎片化的问题，模型上下文协议（model context protocol, MCP）提出一种通用的开放标准，通过标准化模型与外部数据源及工具的连接方式，实现智能体与异构系统间的无缝互操作，降低了对接成本。此外，SWE-agent^[25]通过重新设计智能体-计算机界面，证明适配模型能力的工具界面能有效减少交互噪声；Voyager^[26]则展示了智能体通过维护可执行技能库实现技能自主演化与终身学习的潜力。

除了调用功能型API，智能体对外部知识库工具的动态利用同样是其完成任务的关键。RAG^[27]确立了将预训练参数

化记忆与外部维基百科等非参数化记忆相结合的基础架构，为智能体获取外部证据提供了基石。为了克服被动检索的局限，FLARE^[28]引入了前瞻性主动检索机制，通过预测未来语句并识别低置信度标记来动态触发工具调用，实现了检索时机的自主控制。Self-RAG^[29]进一步构建了闭环反思机制，对检索结果的相关性、支持度及生成质量进行细粒度自我质检，赋予了智能体批判性使用知识工具的能力。针对涉及全库信息的复杂感悟任务，GraphRAG^[30]将检索行为升级为一种基于图索引的规划问题，通过构建实体知识图谱及层级化的社区摘要，实现了对大规模互联知识的深度汇总与全局感知。

5 AI Agent 驱动的意图网络系统

5.1 业务需求

意图网络试图通过自然语言实现配置闭环，是网络自治演进中的关键方向。但从工程落地角度看，目前仍存在三个主要问题：传统自动化缺乏认知弹性，导致语义理解上的偏差；大模型基于概率生成内容，容易产生事实性幻觉；现有系统缺少有效的反馈验证机制，配置过程基本处于开环状态。针对这些问题，本系统的具体业务需求分析如下：

（1）语义理解与实体映射需求

在实际运维中，意图的表述与实体设备指令存在明显差距。意图网络系统需要具备较强的语义理解和实体映射能力，以便将模糊的意图转换为结构化的配置指令。

（2）确定性执行与幻觉抑制需求

大模型存在幻觉，容易虚构一些不存在的参数，例如IP地址或物理端口。这与网络运维对零差错的要求是冲突的。这就要求意图网络系统需要引入检索增强和外部事实锚定机制。

（3）闭环验证与安全防御需求

意图网络系统需要进行安全防御和验证，也就是需要构建前置拦截与后置校验。其中，前置拦截用于实时阻断未经授权的高危操作；而后置校验需要通过动态比较变更前后的流量、丢包率等核心指标，确认策略是否真正生效。

5.2 系统设计与实现

本文设计并实现了一种由AI Agent驱动的神符号协同意图网络架构。该架构遵循功能分层与认知协同两个维度，如图2所示。在结构维度上，系统参照意图网络的标准参考模型，自上而下分为应用层、意图层和网络层。应用层负责意图交互与传递，意图层承担意图解析与参数锚定，网络层专注于指令的原子化执行与物理防护。在认知维度上，系统结合了人工智能的两大流派，形成概率感知与确定执行的互补机制。感知侧采用连接主义范式，利用大语言模型处理自然语言的模糊性；控制侧采用符号主义范式，使用标准作业

程序实现确定性的执行流水线。为验证该架构，本文搭建了完整的原型实验环境：底层使用 Mininet 与 bmv2 构建支持细粒度控制的数据平面，控制面部署自研 SDN 控制器，中间件集成多种异构数据库与消息队列，认知中枢接入 DeepSeek 大模型。

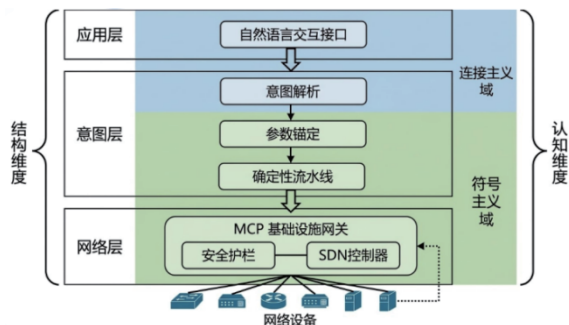


图 2 系统架构

（1）意图解析与双重锚定机制

意图层作为系统的认知中枢，需要解决语义理解偏差的问题。为避免大模型在长链条推理中产生事实性幻觉，本系统在工程实现上采用分段式处理与空间约束机制。

受限语义提取：以“限制用户 A 的带宽为 10 Mbit/s”为例，大模型输出的是结构化的 JSON 文本，里面的信息有意图标签、非结构化实体描述和数值参数。

双重锚定机制：先是进行语义锚定，也就是查询出置信度最高的用户标识符。再进行物理锚定，也就是系统用该标识符进行查询，补全该用户对应的物理 IP 地址与端口等参数信息。

（2）MCP 网关与多级安全护栏

意图网络系统实现了 MCP 网关，这个网关不是简单的指令透传，而是针对每个调用都进行两级校验。并且只有当这两级护栏都通过后，MCP 网关才会调用 SDN 控制器 API 接口。

一级语法护栏（静态强类型校验）：MCP 网关对输入参数进行静态检查，以避免不合理的请求。例如，验证 IP 地址格式是否正确，以及检查数值是否在合法区间内等。

二级策略护栏（动态合规校验）：MCP 网关查询本地安全策略库，如果目标实体被标记为 VIP 客户，又或者当前处于网络变更封禁窗口期，就会抛出合规异常并阻断操作。

（3）遥测闭环与端到端工作流验证

以“限制用户 A 的带宽为 10 Mbit/s”为例，说明端到端的数据流转。底层数据平面周期性地采集对应端口的实时流量统计信息，通过推模式将遥测数据发送到指定 Topic。

MCP Server 则使用 AMQP 协议异步拉取该数据流。

在策略下发后，系统会持续比对配置前后的端口速率。如采集到的端口速率收敛到 10 Mbit/s 的合理误差范围内，系统判定策略生效，并向应用层返回任务成功的确认日志；如

果端口速率仍维持高位，系统会自动触发状态机回滚机制，撤销流表规则，并向应用层发出告警。

6 实验

6.1 实验设计

为验证神经符号协同架构在网络配置生成任务中的有效性，本节实验评估聚焦于该架构的核心创新点：基于思维链的意图解析精度与基于双重锚定的物理参数幻觉抑制效能。架构中的 MCP 网关、多级安全护栏及遥测闭环主要承担确定性逻辑控制与消息路由功能，其实现遵循标准工程规范，属于系统完整性的必要组成部分，但其行为不涉及非确定性推理。因此，本文的量化评估重点考察系统在处理非确定性生成任务时的可靠性，以验证所提神经-符号协同机制的核心优势。

针对当前缺乏领域标准化基准的现状，本文构建了专属评测数据集，具体包含 300 个涵盖显式、口语化与抽象指令的多粒度意图解析用例，以及 100 个高强度幻觉评估用例。在量化评估体系构建上，意图解析阶段采用要求意图分类标签与实体参数实现双重命中的严格完全匹配准确率；配置生成阶段则分别引入表征脱离真实资产库覆盖参数占比的物理参数幻觉率，以及统计必需参数包含占比的配置完整率，从而对模型的事实正确性与实际可用性展开严密论证。

6.2 意图解析能力验证

为全面评估大模型的意图解析效能与系统鲁棒性，本实验设计了基础提示与思维链提示两种对比范式，并在包含 300 例样本的数据集上进行了 30 次独立重复实验。表 1 统计结果表明，在基础提示范式下，系统的整体解析准确率维持在 69.91%，暴露出大模型在直接跨越自然语言到严苛结构化输出时，存在显著且固有的解析瓶颈。相比之下，引入思维链提示范式后，系统的整体准确率获得显著提升，稳定分布在 98.67%~100% 的高位区间。此外，30 次独立实验的准确率标准差仅为 0.31，展现出极高的性能一致性。这一结果证实，思维链提示范式通过强制模型执行显式的分步推理，有效降低了意图转化的认知负荷，并极大地增强了模型对结构化输出边界的约束能力。

表 1 意图解析统计结果

提示范式	单位：%			
	平均值	标准差	最小值	最大值
基础提示范式	69.91	0.30	69.33	70.67
思维链提示范式	99.46	0.31	98.67	100

图 3 的细粒度评测结果深刻揭示了不同提示范式下的解析差异。在基础提示范式下，参数密集的显式指令准确率为 67%，这显著低于参数稀疏的抽象指令的 80%，表明模型在多槽位提取的自回归生成中易发生注意力分散，导致漏提并

产生误差累积。引入思维链提示范式后，显式与口语化指令的准确率均跃升至 100%，证明显式的分步逻辑推演能有效锚定模型的注意力焦点，大幅增强复杂指令的结构化依从性。此外，该范式下全系统仅存的微量误差均集中于抽象指令的动作极性判定上，这说明在缺乏多轮对话上下文的单轮交互中，意图解析性能已逼近理论上限，自然语言固有的语用歧义性成为精度进一步突破的核心瓶颈。

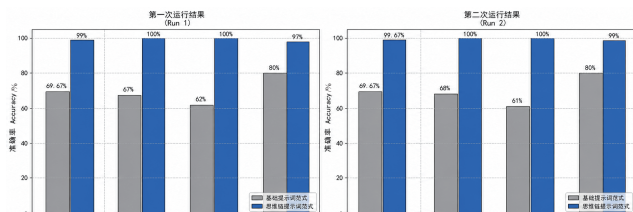


图3 细粒度评测结果

6.3 双重锚定机制消融实验

针对大语言模型生成网络配置时易虚构诸如未分配 IP 等物理参数的幻觉问题，本节依托前述 100 例高强度幻觉评估验证集，通过消融实验定量评估双重锚定机制的抑制效能。实验设置三组渐进对照：直接生成配置的基线组（Group A）无任何外部知识约束；引入强约束提示的对照组（Group B）仅通过提示词显式禁止虚构参数；部署双重锚定机制的实验组（Group C）则依次经由向量匹配与关系查询实现语义与物理的精准约束。系统评估以幻觉率和配置完整率为核心指标。在评估设置上，区别于意图解析阶段 30 次的独立重复实验，本实验引入的确定性外部硬约束极大压缩了模型生成的随机方差。因此，各对照组独立运行 5 次取均值即可获得高度收敛的统计特征，充分确保了量化评估的严谨性。

图 4 双 Y 轴结果揭示了大模型幻觉抑制与配置完整性之间存在显著的安全性性与可用性权衡。

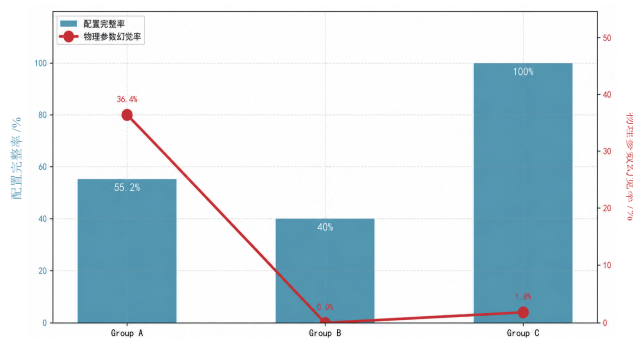


图4 双重锚定机制消融实验结果

直接生成配置的基线组（Group A）幻觉率高达 36.4%，频发虚构 IP 等事实性错误，凸显模型在缺乏知识锚定的多槽位生成任务中的固有局限；而配置完整率为 55.2%，意味着近半数配置缺少必需参数。引入强约束提示的对照组（Group B）虽将幻觉率降至 0%，但配置完整率骤降至 40%，模型因

触发安全规避策略省略关键字段，陷入安全却不可用的退化困境。反观部署双重锚定机制的实验组（Group C）实现全面突破，其配置完整率稳居 100%，且幻觉率收敛至 1.8%，微量误差仅源于广播地址误判。该机制成功驱动模型完成从参数记忆转向知识调用的范式跃迁，从生成源头有效抑制幻觉的产生。

7 结语

本文分别探讨了意图网络、软件定义网络、大模型和 AI Agent，设计并实现了一个由 AI Agent 驱动的神符协同系统。意图解析能力验证和双重锚定机制消融实验表明，该系统表现出较好的意图解析能力和确定性执行能力。在本文已验证的单体智能体基础上，后续工作将探索多智能体协同机制，包括构建多个专用智能体的分工协作体系，研究超大规模网络环境下的任务分解、冲突消解与分布式决策等问题，以应对下一代智能网络管理中复杂配置生成、跨域策略协同和动态自治治理等挑战。

参考文献：

- [1]CASADO M, FREEDMAN M J, PETTIT J, et al. ETHANE: taking control of the enterprise[J]. ACM sigcomm computer communication review, 2007, 37(4): 1-12.
- [2] BEHRINGER M H, PRITIKIN M, BJARNASON S, et al. Autonomic networking: definitions and design goals[EB/OL]. (2018-12-20)[2026-04-17].<https://datatracker.ietf.org/doc/rfc7575/>.
- [3]CLEMM A, CIAVAGLIA L, GRANVILLE L Z, et al. RFC 9315: intent-based networking: concepts and definitions[S]. Internet Engineering Task Force, 2022.
- [4] JEONG J P, AHN Y, GU M, et al. Intent-based network management automation in 5G networks: draft-jeong-nmrg-ibn-network-management-automation-06[S]. IETF, 2025.
- [5] LI F L, FAN G Y, WANG X W, et al. State-of-the-art survey of intent-based networking[J]. Journal of Software, 2020, 31(8): 2574-2587.
- [6] PAN T, YU N B, JIA C H, et al. Sailfish: accelerating cloud-scale multi-tenant multi-service gateways with programmable switches[C]//SIGCOMM'21: Proceedings of the 2021 ACM SIGCOMM 2021 Conference. New York: ACM, 2021:194-206.
- [7]LYU B, SONG E G, PAN T, et al. POSEIDON: a consolidated virtual network controller that manages millions of tenants via config tree[EB/OL]. [2026-04-17]. <https://www.usenix.org/conference/nsdi24/presentation/lyu>.

- [8]PAN T, LIU K, WEI X L, et al.LuoShen:a hyper-converged programmable gateway for multi-tenant multi-service edge clouds[EB/OL].[2026-04-17].<https://www.usenix.org/conference/nsdi24/presentation/pan>.
- [9]VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[PP/OL].(2023-08-02)[2026-04-17].<https://doi.org/10.48550/arXiv.1706.03762>.
- [10]DEVLIN J, CHANG M W, LEE K,et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Brussels: ACL, 2019. 4171-4186.
- [11]BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[C]//NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020:1877-1901.
- [12]TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[PP/OL].(2023-02-27)[2026-04-27].<https://doi.org/10.48550/arXiv.2302.13971>.
- [13]OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[PP/OL]. (2022-03-04)[2026-04-27].<https://doi.org/10.48550/arXiv.2203.02155>.
- [14]GU A, DAO T. Mamba: linear-time sequence modeling with selective state spaces[PP/OL].(2024-05-31)[2026-04-27].<https://doi.org/10.48550/arXiv.2312.00752>.
- [15]WEI J, WANG X Z, SCHUURMANS D,et al. Chain-of-thought prompting elicits reasoning in large language models[C]//NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM,2022: 24824-24837.
- [16]ZHOU D. Learning to reason with LLMs[EB/OL].[2026-04-27].<https://dennyzhou.github.io/LLM-Reasoning-Stanford-CS-25.pdf>.
- [17]DeepSeek-AI.DeepSeek-V3 technical report[PP/OL].[2026-04-27]. <https://arxiv.org/abs/2412.19437>.
- [18] DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[PP/OL].[2026-04-27].<https://arxiv.org/abs/2501.12948>.
- [19]YAO S Y, ZHAO J, YU D, et al.ReAct: synergizing reasoning and acting in language models[PP/OL].(2023-03-10)[2026-04-27].<https://doi.org/10.48550/arXiv.2210.03629>.
- [20]SHINN N, CASSANO F, BERMAN E, et al.Reflexion: language agents with verbal reinforcement learning[PP/OL].(2023-10-10)[2026-04-27].<https://doi.org/10.48550/arXiv.2303.11366>.
- [21]ZELIKMAN E, HARIK G, SHAO Y J, et al.Quiet-STaR: language models can teach themselves to think before speaking[PP/OL].(2024-03-18)[2026-04-27]<https://doi.org/10.48550/arXiv.2403.09629>.
- [22]SCHICK T, DWIVEDI-YU J, DESSÌ R, et al.Toolformer: language models can teach themselves to use tools[PP/OL].(2023-02-09)[2026-04-27].<https://doi.org/10.48550/arXiv.2302.04761>.
- [23]PATIL S G, ZHANG T J, WANG X, et al. Gorilla: large language model connected with massive APIs[PP/OL]. (2023-05-24)[2026-04-27].<https://doi.org/10.48550/arXiv.2305.15334>.
- [24]QIN Y J, LIANG S H, YE Y N, et al.ToolLLM: facilitating large language models to master 16000+ real-world APIs[PP/OL].(2023-10-03)[2026-04-27].<https://doi.org/10.48550/arXiv.2307.16789>.
- [25]YANG J, JIMENEZ C E, WETTIG A, et al.SWE-agent: agent-computer interfaces enable automated software engineering[PP/OL].(2024-11-11)[2026-04-27].<https://doi.org/10.48550/arXiv.2405.15793>.
- [26]WANG G Z, XIE Y Q, JIANG Y F, et al.Voyager: an open-ended embodied agent with large language models[PP/OL].(2023-10-19)[2026-04-27].<https://doi.org/10.48550/arXiv.2305.16291>.
- [27]LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[PP/OL]. (2021-04-12)[2026-04-27].<https://doi.org/10.48550/arXiv.2005.11401>.
- [28]JIANG Z B, XU F F, GAO L Y, et al.Active retrieval augmented generation[PP/OL].(2023-10-22)[2026-04-27].<https://doi.org/10.48550/arXiv.2305.06983>.
- [29]ASAI A, WU Z Q, WANG Y Z, et al.Self-RAG: learning to retrieve, generate, and critique through self-reflection[PP/OL].(2023-10-17)[2026-04-27].<https://doi.org/10.48550/arXiv.2310.11511>.
- [30]EDGE D, TRINH H, CHENG N, et al.From Local to global: a graph RAG approach to query-focused summarization [PP/OL].(2025-02-19)[2026-04-27].<https://doi.org/10.48550/arXiv.2404.16130>.

【作者简介】

周洪利(1981—),男,山东临沂人,硕士,高级工程师,研究方向:软件定义网络、云数据中心、智能网络。

(收稿日期:2026-03-16 修回日期:2026-05-06)