



研究与开发

基于双通道理论的通信认知增强技术研究

廉霄兴, 宋勇, 朱军, 王淑玲, 叶晓舟, 欧阳晔
(亚信科技(中国)有限公司, 北京 100193)

摘要: 结合当前大语言模型增强技术研究进展, 提出了基于双通道理论的通信认知增强技术框架 (telecom augmented cognition architecture, TACA), 解决大语言模型在通信领域应用中专业知识不足、推理能力有限等问题, 实现面向通信复杂问题的认知推理增强、知识融合增强、模型工具增强和自主决策执行增强等应用能力。针对不同的通信业务场景给出了相应技术应用参考, 并通过实验初步验证认知增强方案在解决复杂问题方面的有效性和可行性。因此, 所提出的通信认知增强技术框架可作为未来通信行业大语言模型应用的参考框架。

关键词: 大语言模型; 认知增强; 双通道理论; 自主决策执行; 电信网络

中图分类号: TP181

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2024007

Research on dual process theory based telecom augmented cognition

LIAN Xiaoxing, SONG Yong, ZHU Jun, WANG Shuling, YE Xiaozhou, OUYANG Ye
AsiaInfo Technologies (China) Co., Ltd., Beijing 100193, China

Abstract: Incorporating the current works in augmented large language models (LLM), a dual process theory based telecom augmented cognition architecture (TACA) was proposed. This architecture addressed challenges faced by LLM in the field of telecom, such as lack of domain-specific knowledge and limited reasoning capabilities. It aims to enhance cognitive reasoning, knowledge integration, tooling, and autonomous decision-execution capabilities for complex telco-scenarios. Technical application references for different telco-scenarios were provided, and the effectiveness and feasibility of the augmented cognition solution were preliminarily validated in addressing complex problems through experiments. The proposed TACA can serve as a future reference architecture for the application of LLMs in the telecom industry.

Key words: large language model, augmented cognition, dual process theory, autonomous decision and execution, telecommunication network

0 引言

以 ChatGPT、GPT-4 为典型的大模型 (大语言模型, large language model, LLM) 引发公众对

人工智能技术关注的新浪潮^[1], 并对生成式人工智能未来引起的众多技术应用产生巨大期望。相比过往人工智能技术, GPT-4 在认知推理、泛化性能等方面实现了质的飞跃, 有较强的通用认知能

收稿日期: 2023-08-31; 修回日期: 2024-01-10



力,但是在行业领域应用中,大模型认知能力仍有欠缺,存在领域知识性不足等问题。通信领域涉及的专业知识密集、规则复杂,相关标准准则动态变化,具有实时性、鲁棒性和复杂性等多种要求挑战,以 GPT-4 为代表的通用大模型在实际应用中遇到困难,需要结合行业领域知识和认知能力实现定向增强,以适应通信领域各类复杂任务和认知理解能力需求。本文从人类认知双通道理论出发,提出增强大模型认知能力的技术架构,以全面提升大模型在通信复杂场景的应用能力。

1 通信认知增强相关技术研究

1.1 以 GPT-4 为代表的大模型认知能力实现突破

人类认知能力是指通过思维、经验和感官获得知识和推理的心智过程,涵盖了感知、注意力、思维推理过程,包括长时记忆和工作短时记忆、判断评估、逻辑推理、解决问题、分析决策等多种能力^[2]。“认知智能”是人工智能技术发展的高级阶段,赋予机器数据理解、知识表达、逻辑推理、自主决策执行等能力。在 ChatGPT 和 GPT-4 出现之前,传统机器智能在认知能力方面有一定的局限性。如以知识图谱和符号规则为代表的专家系统和知识工程,该方法具有良好的可解释性,但在面向处理模糊性知识和海量可扩展性方面较差;深度学习出现之后围绕特定任务构建的专业模型,该方法能够实现特定任务如词性分析、文本分类、内容抽取等,然而在知识表征和理解、逻辑推理、泛化性方面仍然非常有限^[3]。认知智能发展路线如图 1 所示。

自 2017 年谷歌推出 Transformers 架构以来^[4],以 Transformers 架构为核心的大模型飞速发展,OpenAI 提出的 GPT 系列模型属于大模型分支,展现出强大认知和泛化能力。而 ChatGPT 和 GPT-4 大语言模型的出现,代表着认知智能出现了重大突破^[5]。GPT-4 能够通过上下文和思维链推理过程来提高认知能力,使其在逻辑推理、指令泛化、上下文理解、阅读理解、常识运用等方面的认知能力有明显提升。通过情景学习和思维链推理, GPT-4 能够理解并连接任何主题,跨模式和跨学科组合不同技能; GPT-4 能够进行知识密集型的任务处理,如编码、数学、物理的证明推理,在法律和医学专业测试中超出人类水平表现;还可以使用工具与外部进行迭代交互,展现出计划和解决复杂问题能力。GPT-4 也表现出优越的多模态认知理解能力,根据图像特征进行高阶的语义分析和逻辑推理能力。这表明以 GPT-4 为代表的大模型具备了较强认知能力,能够从海量数据和人类指令中学习到世界知识和部分专业知识,并能根据所学知识进行分析、评估、推理和决策等灵活运用。尽管 GPT-4 认知能力取得了突破,但仍面临着事实性错误、幻觉问题、缺乏计划性、无法全局视角和推理能力有限等诸多问题,其本质仍然属于统计模型范畴,对于符号规则概念的理解运用仍然有限,这限制了其在专业知识密集、任务策略复杂以及多步推理决策场景中应用。

图 1 展示了认知智能的发展路线。从 1990 年的专家系统开始，经过 2010 年的深度学习，最终发展到 2018 年的大模型。专家系统基于符号和逻辑符号规则系统，对知识进行人工定义，基础任务包括分类、回归、逻辑、聚类，典型算法有逻辑回归、符号规则和一阶谓词。深度学习基于更多数据和更大算力，训练深度神经网络模型，释放更大能力，局部任务包括词性分析、文本分类、内容抽取，典型算法有 CNN、LSTM 和 GAN。大模型基于大规模数据和强算力，具有强大认知和泛化能力，向多模态、具身、科学等方向延伸，初步通用智能包括上下提示、情景学习、思维链推理，典型模型有 Transformer、GPT-4 和 ChatGPT。

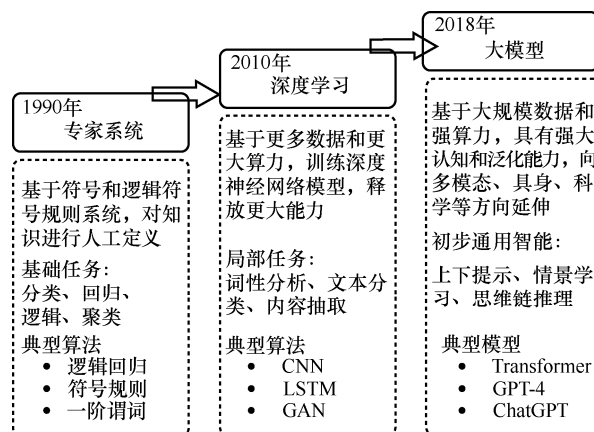


图 1 认知智能发展路线

1.2 基于双通道理论的通信认知增强技术进展

随着大模型在行业领域应用热潮到来,越来越多学者尝试基于双通道理论(dual process theory, DPT)架构来增强大模型认知能力。双通道理论最早由心理学家 Jonathan^[6]提出并建立基本

理论体系, 经过 Kahneman^[7] 认知科学家们的不断发展和演进, 已成为在研究人类认知、决策、思考推理领域的重要参考模型。双通道理论认为人类解决问题的思维过程分为两套系统, 分别为系统 1 (启发式系统, heuristic system) 与系统 2 (分析式系统, analytic system), 并依靠短时记忆和长时记忆进行支撑。系统 1 是启发直觉式过程, 通常是快速的、无意识的、自动化的、联结隐式的思考方式; 系统 2 是思辨分析式过程, 是缓慢的、逻辑思维、回溯反思和有意识 (基于复杂认知结构) 的显式推理过程, 能够基于认知模型整合外部信息进行综合判断决策。人类通过两种思维模式的混合与协作处理复杂任务。

在基于双通道理论的大模型认知增强研究方面, OpenAI^[8] 认为 GPT-4 等大模型可类比于系统 1, 其推理过程来源于自回归式的概率预测, 能够快速生成与问题相关的输出, 但缺乏足够逻辑性和可解释性, 这与人类系统 1 思考模式类似。图灵奖获得者 Yann^[9] 认为自回归式的大模型也可类比系统 1, 在认知能力方面有部分提升但仍面临许多重大问题, 需要通过构建世界模型认知体系增强复杂推理和认知能力。业界对基于双通道理论的认知增强技术进行探索并取得进展, 包含符号逻辑规则结合、知识图谱/知识库融合、基于思维链的复杂认知架构、自主决策执行代理等技术路线。如 Maxwell 等^[10] 将 GPT-3 (可类比系统 1) 与逻辑推理模块 (可类比系统 2) 相结合, 改进文本生成的一致性和合理性, 解决了 GPT-3 违反常识和逻辑约束的问题。知识融合方面, 大模型可以通过向量检索与外部知识库和信息实时连接, 有效减少模型幻觉问题; 如交替性检索与思维链 (interleaving retrieval with chain-of-thought, IRCot)^[11] 推理构建了动态的信息检索机制, 将检索到的信息作为上下文线索融入多步思维链推理过程中, 增强了信息检索精准性和复杂推理能力。在构建复杂认知结构方面, 思维树 (tree of thoughts, ToT)^[12] 在系统 2 中构建了一个

树状认知结构, 通过构建和维护认知结构进行反思、分析和判断, 实现思辨式的思考推理过程; 树搜索语言代理 (language agent tree search, LATS)^[13] 将蒙特卡洛树搜索中模拟和回溯过程以自然语言理解和评估的形式进行操作, 实现了基于显式语义结构的自主探索和回溯, 大幅增强复杂推理能力; 类似工作还有多链元推理 (multi-chain reasoning, MCR)^[14]、思维图 (graph of thoughts, GoT)^[15] 等。在自主决策执行层面, HuggingGPT^[16]、OpenAGI 等^[17] 以自主代理形式将大模型思维规划过程和执行功能结合, 对流程执行策略、思维链提示、外部插件接口、向量检索系统进行组合编排, 实现自动调用多模态模型和操作科学工具仪器, 扩展了面向复杂场景的决策执行能力。以上研究虽然方法和技术路线不同, 但均受双通道理论启发增强了面向复杂问题的推理和分析决策能力。

1.3 电信网络智能化领域认知需求分析

随着 5G 网络高体系性、高复杂性、超大规模、高可靠性的特点, 以及未来 6G 网络系统将承载支撑更加个性化、泛在化、情景化的服务, 电信运营商在业务体验及网络运营领域有非常强烈的认知智能需求。在其网络全生命周期管理规划中, 贯穿感知、分析、预测、决策和控制的认知闭环构成了自智网络智能化的核心。同时, 欧洲电信标准组织 (European Telecommunications Standards Institute, ETSI)、第三代合作伙伴计划 (3rd Generation Partnership Project, 3GPP)、电信管理论坛 (TeleManagement Forum, TMF) 等标准化组织提出了支持网络智能化定制的体系化架构, 其中通信网络的认知能力成为各标准组织网络智能化体系架构的核心。如 ETSI 下属零接触网络服务管理行业规范组 (ISG Zero touch network and Service Management, ISG ZSM)^[18]、通用自主网络架构 (generic autonomic network architecture, GANA)^[19] 等框架中对于通信知识理解、推理分析、策略执行、归纳总结和控制闭环等进行了详细设计要求;



TMF^[20]从运营和闭环角度提出自智网络三层四闭环核心框架。

在大模型与通信领域结合方面,刘贤松等^[21]提出了构建通信行业领域大模型的技术方案,并提出结合网络知识图谱/知识库提升大模型事实准确性等技术方向;毛志伟等^[22]系统分析了当下通信行业和大模型结合路径,阐释了GPT生成式大模型在5G基站规划设计、意图驱动网络设计、智能网络规划、智能决策支持等网络全生命周期管理应用前景;李涛等^[23]系统梳理了面向高阶演进的网络智能化框架,重点阐述了感知/认知/决策的智能闭环流程和相应功能模块作用,推动电信运营商网络敏捷高效和智能。综上,业界在网络智能化框架和标准规范中对认知能力提出了较高要求,然而仅停留在概念模型层面,难以实现体系化落地实施;业界在通信网络领域引入大模型提出了初步实施方案,然而对于如何增强大模型在通信领域专业性和知识性这一核心问题,并未有体系化论述和框架设计。

通信领域知识范畴较广、规则复杂,仅依靠通用大模型难以胜任通信业务。因而十分有必要借鉴双通道理论,解决大模型固有局限和专业能力不足问题。本文结合当前大模型前沿技术趋势进展,系统梳理双通道认知增强技术框架,阐述基于双通道理论的通信认知增强技术框架和体系内涵,构建面向通信行业的大模型知识融合增强、分析推理增强、模型工具增强和自主决策增强能力,为面向高阶的自智网络演进提供技术路线和实施参考。

2 基于双通道理论的通信认知增强技术框架

2.1 设计思路

系统1与系统2组成:系统1由参数化的人工智能模型构成,通过无监督训练、监督微调和人机对齐和强化学习等方式训练,存储了参数化的隐式知识,其推理过程本质是统计概率输出;

系统2由显式结构化认知体系组成,包含了规划、分析、监控、验证、评估等高阶认知提示策略,并通过执行框架调动集成外部知识信息、记忆系统、逻辑规则和工具插件模块,经过显式多步的思维推理过程最终实现整合输出。

双通道间交互转化:系统2在认知提示策略指引下结合知识信息、逻辑规则等,构建上下文提示输入给系统1,系统1根据上下文提示的任务引导、知识线索和信息反馈,进行直觉式思考推理输出,系统2将系统1的输出与工具检索、逻辑规则、记忆系统等进行信息整合并构建新的上下文提示,经过双通道间频繁交互转化,实现思辨式反复思考。

多步动态推理过程:相比系统1的直觉式推理输出,系统2执行基于思维结构的多步动态推理过程,在任务特定的认知提示策略引导下进行规划、分解、评估、分析和探索等复杂思考,每步推理思考过程中均可融合知识信息后迭代优化,通过多步推理思考解决复杂任务。

可修正、可解释:由于系统2由显式化的认知体系构成,其思维推理过程和模型输出均为语义输出,执行过程可解释,同时具备可优化、可修正潜力。

2.2 构建基于双通道理论的通信认知增强技术框架

构建基于双通道理论的通信认知增强技术框架(telecom augmented cognition architecture, TACA),包含基于大模型的系统1、融合通信认知体系的系统2以及记忆系统底座,该框架借鉴人类处理复杂问题时的认知架构,融合规划、分析、监控、验证、评估等思维链推理能力,有效利用外部信息、知识、工具和决策模块,实现基于意图和任务场景的自主决策、反思和执行闭环能力。框架包含基础设施层、认知增强层和交互感知层。

(1) 基础设施层:为大模型提供基础算力、计算基础框架和网络通信服务,包含大规模分布

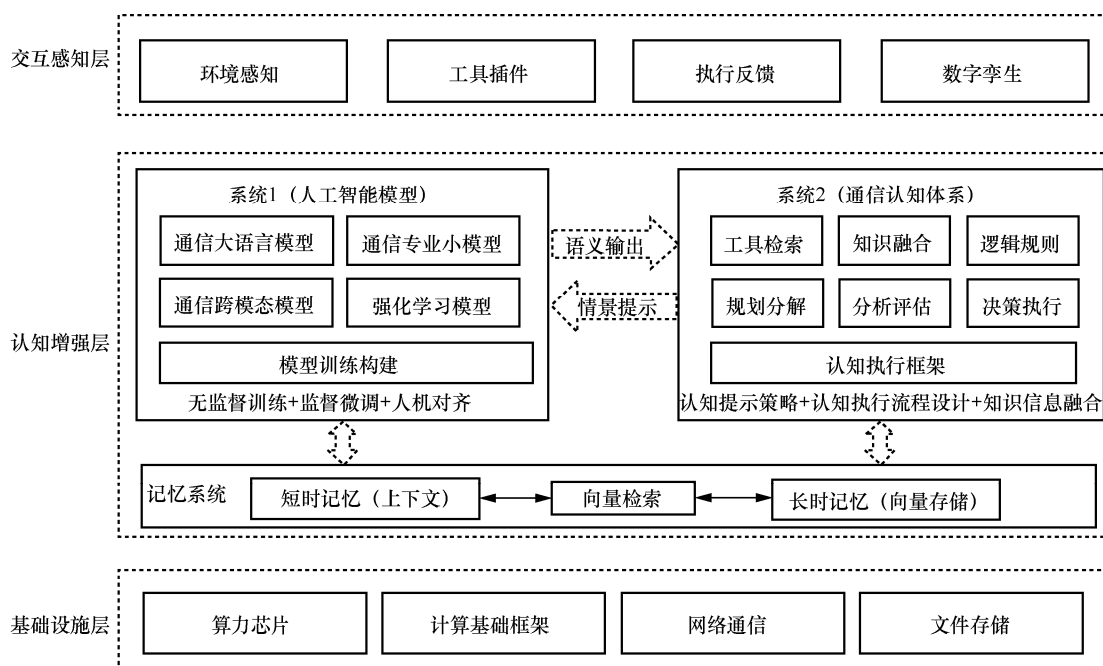


图2 基于双通道理论的通信认知增强技术框架

式训练、底层资源弹性调度、高效计算通信以及提供训练推理框架等基础设施，也包含支撑模型训练和认知推理的数据与知识存储。

（2）认知增强层：实现通信认知增强的核心层，包含抽象化的系统1、系统2以及记忆系统三大系统。

系统1：由融合通信行业领域知识训练构建的人工智能模型组成，包含通信大语言模型、跨模态模型、专业小模型和强化学习模型等类型。大语言模型和跨模态模型可以接受系统2的认知提示策略指令，系统1本身包含了参数化知识，具备情景学习和思维链推理能力，在上下文提示的任务引导、知识线索和信息输入下，可进行直觉式思考推理输出，在准确性、时效性不敏感场景中可直接使用系统1输出。系统1包含了从无监督训练、监督微调到人机对齐的通信人工智能模型训练构建流程。无监督训练主要对通用语料和通信行业标准、法规文件、技术文献等通信领域通用语料进行无/自监督训练；监督微调通过通信专业指令和任务精调，使大模型“听懂”通信领

域指令；人机对齐通过强化学习等方式，优化模型在面向特定任务的输出行为，使模型输出与电信业务规则规范一致。

系统2：由显式化认知提示策略、认知执行流程设计和知识信息融合构成，实质是构建了针对通信业务场景下的通信认知体系和认知执行框架。认知提示策略包含了针对特定场景的规划、分析、监控、验证、评估等提示模板和执行策略。认知执行流程能够整合外部知识、逻辑规则、工具插件模块和记忆系统等多种模块，将通信认知提示策略和外部信息组合起来融入连贯、目标一致的多步动态推理过程，并在多步推理过程中实现知识信息检索融合、规划任务分解、回溯分析评估、模型工具调用、自主决策行动等复杂认知执行能力。系统2依托情景学习和思维链能力，允许外部知识信息融合反馈，提高了专业性和准确性，可解决网络运维故障分析、无线网络规划设计、网络配置方案生成等通信复杂应用。

记忆系统：包含工作记忆（短时记忆）和长时记忆。工作记忆指大模型所能接受的上下文窗



口限制,系统 1 根据工作记忆中的指令、样例示范、背景知识等内容实现情景学习;长时记忆指能够持久化存储的数据库系统,包含过往对话记录、外部结构化信息、非结构化音视频、知识图谱知识库和实时数据等信息;向量检索模块可实现对内外部结构化、非结构化数据信息向量嵌入等功能,供系统 1 需要时进行基于语义的访问查询检索,或输向短时记忆形成上下文。

(3) 交互感知层:承担了外界环境与认知增强决策执行的反馈感知功能,进行环境感知、工具插件调用和执行反馈,接收认知增强层的指令并通过语义和多模态信息反馈给认知增强层,网络数字孪生可针对无线网、核心网和传输网进行仿真与验证。

2.3 通信认知增强技术形式化问题定义

基于双通道的认知增强本质上基于大模型本身启发式思考过程,结合外部知识、认知提示策略、模型工具融合、自主执行交互反馈来提升对复杂问题理解和任务能力提升,通过体系化、结构化和内外部的信息整合,全面提升大模型的认知推理和问题解决能力。通信认知增强问题形式化定义如下:

定义:认知增强映射:

$$O(F, \mathcal{K}, A, \mathcal{R}, \mathcal{M}) : \mathcal{L}_\theta(x, z_1, \dots, z_n) \rightarrow y \quad (1)$$

系统 1: 参数为 θ 的大模型 $\mathcal{L}_\theta$, $\mathcal{L}_\theta(x)$ 代表大模型对问题 x 输出, y 为期望输出。

系统 2: $O(F, \mathcal{K}, A, \mathcal{R}, \mathcal{M})$, 依托认知执行流程设计编排,实现认知能力增强。

(1) F 为认知提示策略集合, $F = \{f_1, f_2, \dots, f_n\}$ 包含针对通信业务场景问题的各类认知提示和流程策略,引导构建结构化认知体系和认知执行流程。

(2) $\mathcal{K} = \{k_1, k_2, \dots, k_n\}$ 为不同领域知识信息检索融合过程。

(3) $A = \{a_1, a_2, \dots, a_n, a_i \in \pi\}$ 为不同的行动策略集合,在规定的空间 π 执行的相应操作行动,

如通信数据库访问、计算器访问、搜索查询、模型联动、函数调用、数据存储等操作,或者通过大模型输出指令操作小模型。

(4) $\mathcal{R} = \{r_1, r_2, \dots, r_n\}$ 为交互感知层输入的环境信息反馈。

(5) $\mathcal{M} = \{\mathcal{M}_s, \mathcal{M}_l\}$, $\mathcal{M}_s$ 为短时记忆上下文窗口,会选择性地对思维提示和外部信息进行记忆存储; $\mathcal{M}_l$ 为长时记忆,对当前历史认知行动, $z_1, \dots, z_i$ 外部知识信息检索 $\mathcal{K}$ 、环境信息反馈 $\mathcal{R}$ 进行向量存储并供检索。

多步推理过程: $z_1, \dots, z_n$, 中间过程 $z_i = \{t_i, f_i, k_i, a_i, r_i\}$, 包含当前过程系统 1 思考输出 t_i 和认知提示、知识融合、执行操作、环境反馈等。每步推理过程 $z_i \rightarrow z_{i+1} : \mathcal{L}_\theta^{\text{prompt}}[f_{i+1}(z_1, \dots, z_i, \mathcal{K}, \mathcal{R}, \mathcal{M})]$ 。即每一步的思维推理过程由认知提示策略结合历史记录、外部环境反馈、知识信息融合共同决定,而思考过程本身是由系统 1 大模型 $\mathcal{L}_\theta$ 来实现,系统 2 在于通过认知提示策略和外部信息构建精细化的语义上下文 $f_{i+1}(z_1, \dots, z_i, \mathcal{K}, \mathcal{R}, \mathcal{M})$, 构建迭代式的规划、评估、决策、反思和行动的复杂认知思维行动过程,为解决复杂、动态和不确定性问题带来了实施操作空间。多步推理过程示意图如图 3 所示。

3 通信认知增强技术内涵及应用参考

3.1 通信复杂认知推理增强

复杂推理能力是指根据已有知识信息进行理解判断分析的过程,一般包含逻辑推理、常识推理、多跳推理、分析推理等多个领域。通信系统专业性较高、耦合因素较多、体系复杂,如无线高清质量分析、5G 数据时延等问题,而大模型专业性和推理能力难以满足要求。通过系统 2 的认知提示策略,如业务规划分解、回溯反思、评估检验、前瞻预测等,并将模型的中间思维过程 z_i 构建为问题认知结构(一般为图、链、树等)存储在记忆系统中,针对复杂认知结构进行分析推理,



构建下一步分析推理所需提示,实现通信复杂组合式探索问题的解决能力。

3.2 通信领域知识融合增强

大模型参数可认为存储了隐式知识,能够灵活泛化应用但具有不确定性、不可解释和难以动态更新等缺点;以符号规则、知识库、知识图谱等存储了显式知识,知识精准、可解释,但在泛化性和理解运用方面有局限。大模型和知识图谱/知识库结合,可发挥不同知识形式优势增强大模型认知能力。

通信业务知识包含策略类、事实类和模块类3种知识形式,策略类知识包含解决问题的思维过程示例或者思考步骤,引导大模型思维链按预定方向进行推理,如5G网络切片策略、资源调度策略等;事实类知识包含实际网络拓扑、运维手册、工单日志、性能指标等基础数据;模块类知识一般指外部知识模块(如网络数字孪生仿真、代码编译器、符号推理等)。通信双通道认知增强技术框架将显式知识和隐式知识结合,通过多步推理、优化提示、检索增强等方式,提升通信领域知识理解与确定性,充分发挥知识价值解决复杂问题。其形式化定义如下:

$$O(f, \mathcal{K}, \mathcal{M}) : \mathcal{L}_\theta(x, z_1, \dots, z_n) \rightarrow y \quad (3)$$

检索式知识增强:通过大模型对通信领域业务知识自动分段、切片,将知识语料向量化,存入通信领域的向量数据库。当用户与大模型交互查询时,会对查询问题进行向量化,并在记忆系统中进行向量相似度检索,然后将检索匹配到答案和线索结合上下文提示返回给大模型。根据检索到的知识内容,结合大模型的整合能力实现知识推理和表达,检索式知识增强可以克服大模型本身的不确定性和幻觉问题,有效提升问题回答的准确性和严谨性,能够针对回答内容回溯消息引证来源,提升可解释性。

动态融合类知识增强:当知识库与问题相差较远时,检索式知识增强对知识理解利用有限,需要将知识融入复杂认知推理过程中,实现知识发现、知识融合理解、知识推理和知识运用的全流程贯通。系统2将检索到的知识整合融入多步思维推理中,在每步思维过程对检索到的知识信息进行学习融合。通信领域知识融合增强示意图如图5所示,在通信投诉处理和故障处置场景中,需要根据检索历史投诉工单、网络运维知识库、网络实时数据和故障处置策略等多种知识来给出

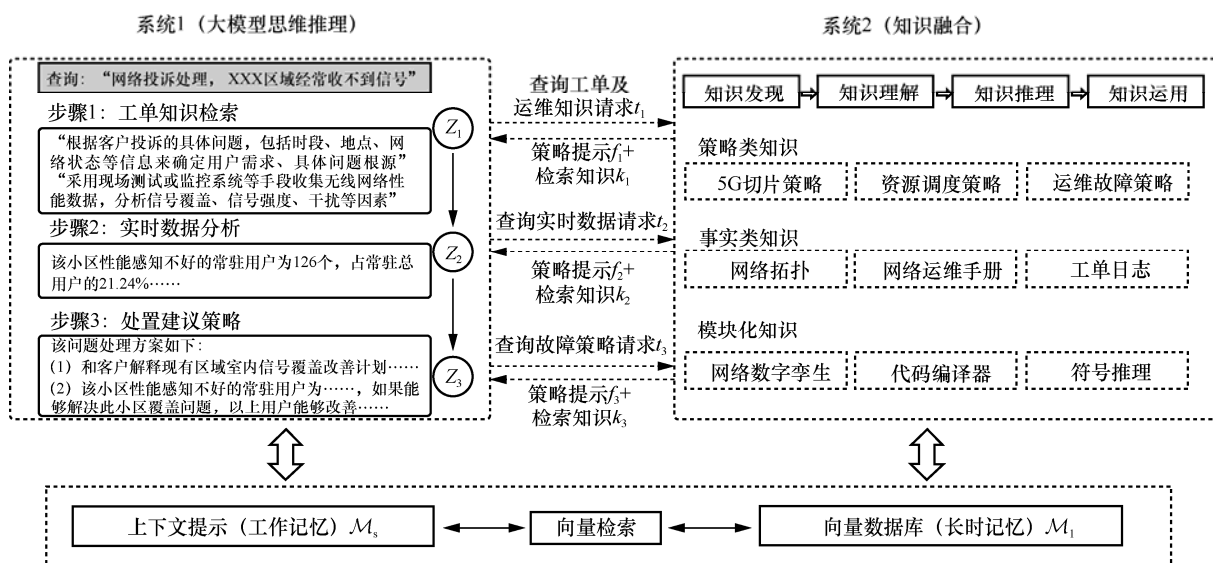


图5 通信领域知识融合增强示意图

最终建议, 每步推理过程中根据提示策略对相应知识库/知识图谱进行检索查询, 并将检索到的知识进行分析总结推理, 此过程动态融合了多源知识信息, 提高通信业务领域知识推理和融合应用能力。

3.3 通信领域模型工具融合增强

通过调用外部工具、操作系统和专业模型进行信息获取、执行操作和任务分配, 可以大幅增强大模型认知、执行和决策能力, 提升大模型专业性和严谨性。依托大模型在认知提示策略下针对任务的意图识别、策略制定、应用程序接口(application program interface, API)调用等能力, 作为连接多个具体任务模型的通用接口, 实质上大模型构成了模型工具调用的规划和策略中心。在针对具体通信任务场景时, 大模型依托系统 2 认知提示策略 F 产生多步思维推理过程, 在每步推理过程中大模型根据情景学习自主判断调用工具插件和执行决策、输入输出参数, 即行动操作 A , 系统 2 工具检索模块将 API 反馈的信息 $\mathcal{R}$, 反馈给系统 1 以进行下一步的认知推理直到问题得到解决。形式化定义如下:

$$O(f, A, \mathcal{R}, \mathcal{M}): \mathcal{L}_\theta(x, z_1, \dots, z_n) \rightarrow y \quad (4)$$

通信专用类 API 一般包含工具插件、专业模

型和配置接口等不同形式。具体如网络运维系统、网络数据平台和专业模型、网络配置操作、网络数字孪生或运维接口等专用 API。通信领域模型工具融合增强示意图如图 6 所示, 在无线网络覆盖优化场景中, 在系统 2 认知提示策略和认知执行流程设计下, 无线网络问题分解成无线网络问题识别、无线问题分析诊断、网络优化方案生成和优化方案仿真验证 4 个推理过程, 分别调用了无线容量问题识别、根因定位、覆盖容量优化专业模型, 并依次将模型工具反馈信息传入下一步推理过程, 最后将优化方案送入网络数字孪生进行验证。系统 2 的工具检索和 API 管理模块保存着 API 能力描述、执行文档和接口标准, 保证了调用可靠性、准确性。基于大模型的认知分析决策, 结合专业模型工具插件能力, 最终实现面向复杂任务的通信认知增强技术。

3.4 通信领域自主决策执行增强

在面向意图驱动网络设计场景中, 工作流程复杂、任务类型多变, 通过构建具有自主决策执行能力的通信自主代理, 实现基于网络意图的网络运维流程的最小单元。通信自主代理可结合任务需求, 自行调用系统工具插件及专业小模型, 并通过组件集成和流程编排能力, 实现网络自规

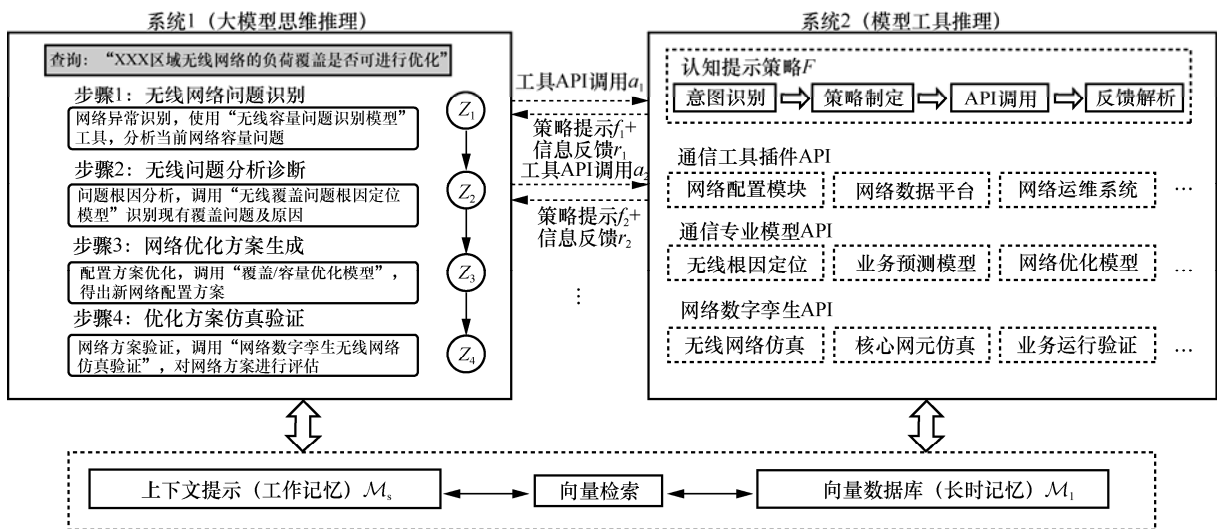


图6 通信领域模型工具融合增强示意图



划、自部署、自优化、自演进。具体实施层面，每个自主代理基于系统 2 的认知提示策略实现不同的场景执行能力，具备独立的工具插件、知识文档和记忆系统，并结合通信业务流程驱动编排设计，完成复杂网络运维流程，实现面向高阶的意图驱动网络演进。基于双通道理论的自主代理架构如图 7 所示，系统 1 为通信大语言/跨模态模型和强化学习模型，系统 2 包含自主代理配置策略、任务执行模块、推理思考模块和知识融合模块。

自主代理配置策略：通过从多个维度描绘代理所承担的职责，为其创造任务背景和属性，同时对代理的能力边界和代理之间协作关系进行设定，使其在实际应用中更真实可信。

任务执行模块：根据目标计划，能够调用通信工具插件 API 进行自动执行指令、操作或决策，可执行诸如信息检索、数据查询、数学计算、复杂语言生成和图像分析等多种动作，并获取环境信息交互反馈。

推理思考模块：结合任务目标、职责属性和网络知识信息实现任务分配、评估决策等工作，将目标分解为子计划和子步骤，结合外部反馈信息和知识信息进行多步分析评估推理等过程。

知识融合模块：与前述通信领域知识融合

相同。

向量数据库记录了过往的操作日志和信息反馈，供通信自主代理检索，改进执行策略和执行任务能力，提高处理复杂问题能力。可通过强化学习方式 O ，不断学习优化执行策略和 workflows，其形式化表达如下：

$$O(f, \mathcal{K}, A, \mathcal{R}, \mathcal{M}) : \mathcal{L}_\theta(x, z_1, \dots, z_n) \rightarrow y \quad (5)$$

4 通信认知增强技术实验验证

面向电信经营数据查询场景搭建原型平台和试点验证，用户提问以自然语言对话形式为主，如“每天的通话新增了多少用户”“最近这个月 5G 网络新增用户数趋势如何”等。由于通信专业术语较强、数据下钻关联结构复杂，大模型直接回答用户查询时无法准确对指标拆解和配置，易出现答非所问和幻觉问题。面对趋势分析、关联分析等洞察分析要求时更难以胜任。

4.1 认知增强方法设计

整体验证思路包含两部分：一是指标检索排序，结合向量检索和 BM25 信息匹配方法提升检索精准性；二是多步动态推理，结合大模型上下文学习、认知提示策略增强复杂推理能力。智能经分场景通信认知增强流程如图 8 所示。

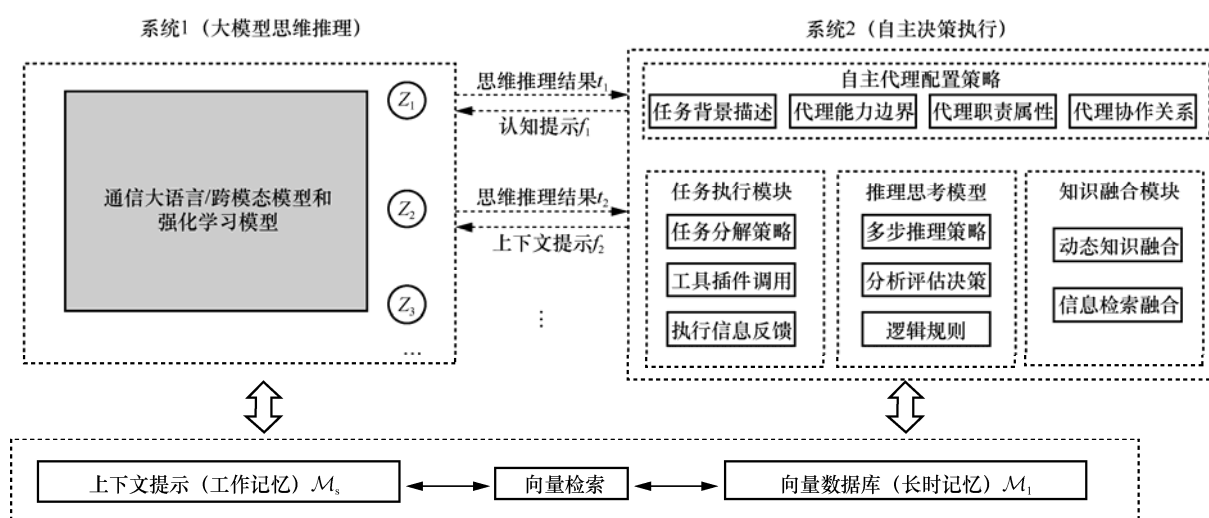


图 7 基于双通道理论的自主代理架构

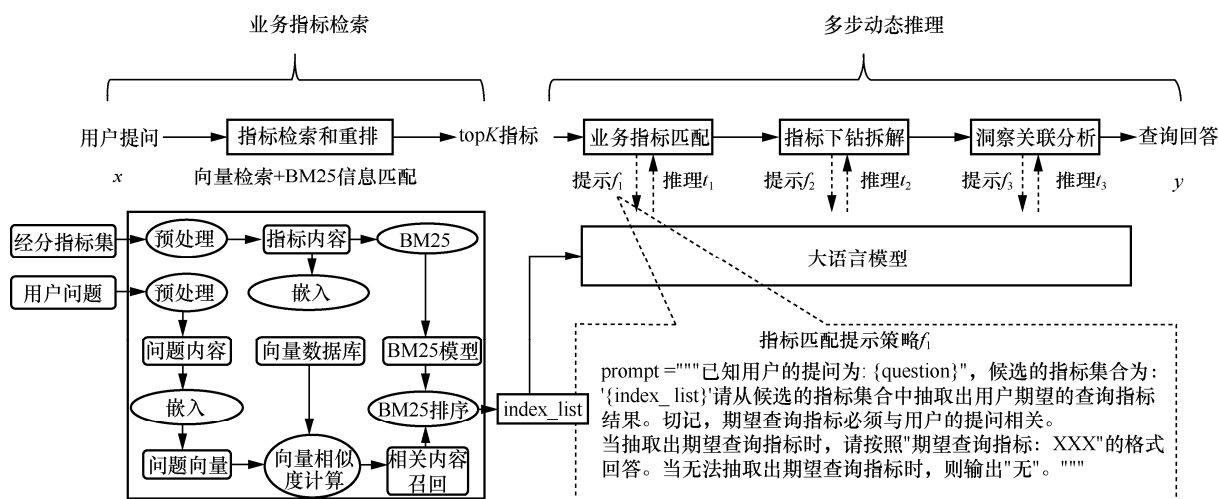


图 8 智能经分场景通信认知增强流程

指标检索排序 (retrival & reranking, R&R) : 采用了 BM25 关键信息匹配结合向量检索方法, 首先根据用户问题和经分指标集向量相似度计算检索出 top20 指标, 后采用 BM25 方法对 top20 指标进行重排。最终从 200 多项数据指标名称中, 找到最相关的 top10 指标, 将候选指标传入大模型上下文提示中, 供后续推理使用。

多步动态推理 (multi-hop reasoning, MHR) : 构建场景化提示工程, 利用大模型逻辑推理+增强分析提示模板, 实现趋势预测、多维度对比关联等数据洞察能力, 生成多维度数据分析输出。

- ① 业务指标匹配: 将用户的提问与 top10 候选指标集拼接, 让大模型判断用户的期望查询指标;
- ② 指标下钻拆解: 从用户的提问中提取出时间、地区和其他维度进行下钻分析;
- ③ 洞察关联分析: 针对客户意图, 根据相应认知提示策略进行趋势、预测、异常分析。

4.2 数据性能验证

验证方法及数据集: 对大模型业务指标检索和多步动态推理进行联合验证, 业务指标检索主要针对 top1、top5、top10、top20 指标召回率进行验证; 多轮动态推理测试采用 ChatGPT、ChatGLM-130B、ChatGLM2-6B、ChatGLM-6B、Vicuna-13B 等多个开源大模型或 API, 16 个数据工程师手工编写 40 份测试用例, 并从指标准确性、分析完整性、分析准确性、逻辑一致性 4 个维度进行主观评分。每个测试用例包含 2~3 轮自然语言追问, 如“那宽带呢”“那杭州呢”“和上月比怎么样”。对所有评委评分进行最小最大 (Min-Max) 归一化处理, 并按技术方案对测试用例进行平均, 结果见表 1、表 2。

检索排序结果表明采用 BM25 结合 GanymedeNil_text2vec-large-chinese 向量嵌入方法的指标召回率最优, 比仅 BM25 重排方法在不同

表 1 业务指标检索排序召回率测试

序号	检索排序	Recall@1	Recall@5	Recall@10	Recall@20
1	BM25+Hanlp 分词检索	39.90%	86.64%	95.16%	97.46%
2	BM25+GanymedeNil_text2vec-large-chinese	40.66%	87.10%	95.61%	97.46%
3	Simcse-chinese-roberta-wwm-ext	35.31%	80.11%	91.61%	93.61%
4	GanymedeNil_text2vec-large-chinese	36.53%	84.8%	93.28%	95.92%
5	Hfl_chinese-roberta-wwm-ext	32.46%	76.98%	90.27%	92.74%



表 2 大模型推理回复评估测试

序号	大模型推理方案	ChatGPT	ChatGLM-130B	Vicuna-13B	ChatGLM-6B	ChatGLM2-6B
1	MHR + R&R(TACA)	0.954	0.872	0.549	0.653	0.764
2	SSR + R&R	0.913	0.840	0.518	0.645	0.749
3	SSR	0.425	0.348	0.205	0.277	0.307

K 值下均有提升。在大模型推理部分,采用多步动态推理结合指标检索排序即 MHR+R&R 的通信认知增强方案取得了最佳结果,相比单步推理(single-step reasoning, SSR)结合检索排序方案, MHR 所带来的性能提升在 ChatGPT 和 ChatGLM-130B 模型较为突出,这表明多步动态推理带来的认知增益在模型规模较大时更加适用,而较小规模模型的多步动态推理提升效果有限。检索排序带来了较大幅度提升,可提高回复精准性并减少幻觉效应,且对所有模型均适用。实验整体验证了通信认知增强技术框架中知识融合增强和认知推理增强,并证明了多步动态推理的有效性和适用范围。

5 结束语

随着大模型在行业领域应用的热潮到来,通信行业迫切需要解决大模型在通信专业知识不足、推理能力有限等问题。本文从人类认知双通道理论出发,分析了通信领域的认知能力需求,提出了基于双通道理论的通信认知增强技术框架。并针对不同的通信业务场景,如 5G 网络时延故障、无线网络覆盖优化、网络投诉处理、意图驱动网络等,给出了基于双通道理论的通信认知增强技术的应用参考,并通过智能经分场景初步验证了所提方案增强认知推理能力的有效性和可行性。

展望未来,基于双通道理论的通信大模型认知增强技术还有以下几方面需进一步探索完善:一是双通道理论技术架构内涵的进一步延伸和明确,系统 1 和系统 2 的模块内涵、协作融合机制需要进一步在工程化实践中探索;二是双通道技术体系与自智网络逻辑架构的融合,双通道架构

如何与控制面、数据面、通信面等逻辑架构之间复杂的协同^[21],实现平台化、一体化的智能服务体系;三是围绕双通道的认知能力进一步提升,如系统 1 结合强化学习提升决策执行的可靠性,系统 1 结合系统 2 和人类反馈信息进行训练微调和人机对齐优化工作,系统 2 如何实现多步骤、多层次认知能力,以及系统 1 和系统 2 联合训练等方向将非常值得进一步探索。

参考文献:

- [1] 张嗣宏,张健.以 ChatGPT 为代表的生成式 AI 对通信行业的影响和应对思考[J].电信科学,2023,39(5):67-75.
ZHANG S H, ZHANG J. Impact and countermeasures of generative AI represented by ChatGPT on the telecom industry[J]. Telecommunications Science, 2023, 39(5): 67-75.
- [2] Lexico. Oxford university press and dictionary.com “cognition”[EB]. 2020.
- [3] 王昊奋,王萌.“神经+符号”:从知识图谱角度看认知推理的发展[J].中国计算机学会通讯,2020,16(8):52-56.
WANG H F, WANG M. Neuro-symbolic: the development of cognitive reasoning from the perspective of knowledge graphs[J]. Communications of the CCF, 2020, 16(8):52-56.
- [4] ASHISH V, NOAM S, NIKI P, et al. Attention is all you need[C]//NIPS2017: Proceedings of the 31st International Conference on Neural Information Processing Systems. California: Curran Associates Inc, 2017: 6000-6010.
- [5] 桑基韬,于剑.从 ChatGPT 看 AI 未来趋势和挑战[J].计算机研究与发展,2023,60(6):1191-1201.
SANG J T, YU J. ChatGPT: a glimpse into AI's future[J]. Journal of Computer Research and Development, 2023, 60(6): 1191-1201.
- [6] JONATHAN S B E. In two minds: dual-process accounts of reasoning[J]. Trends in Cognitive Sciences, 2003, 7 (10), 454-459.
- [7] KAHNEMAN D. Thinking, fast and slow[M]. New York: Farrar, Straus and Giroux, 2011.
- [8] BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of artificial general intelligence: early experiments with GPT-4[EB]. arXiv preprint, 2023, arXiv: 2303.12712.
- [9] YAN N. A path towards autonomous machine intelligence[EB]. 2022.

- [10] MAXWELL N, MICHAEL H T, JOSHUA B, et al. Improving coherence and consistency in neural sequence models with dual-system, neuro-symbolic reasoning[J]. Advances in Neural Information Processing Systems, 2021: 25192-25204.
- [11] HARSH T, NIRANJAN B, TUSHAR K, et al. Interleaving Retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions[C]//ACL2023: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023: 10014-10037.
- [12] YAO S, YU D, ZHAO J, et al. Tree of thoughts: deliberate problem solving with large language models[J]. arXiv preprint arXiv:2305.10601, 2023.
- [13] ZHOU A, YAN K, SHLAPENTOKH-ROTHMAN M, et al. Language agent tree search unifies reasoning acting and planning in language models[J]. arXiv preprint arXiv: 2310. 04406, 2023.
- [14] YORAN O, WOLFSON T, BOGIN B, et al. Answering questions by meta-reasoning over multiple chains of thought[J]. arXiv preprint arXiv: 2304. 13007, 2023.
- [15] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: solving elaborate problems with large language models[J]. arXiv preprint arXiv:2308.09687, 2023.
- [16] SHEN Y L, SONG K T, TAN X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in hugging face[J]. arXiv preprint arXiv: 2303.17580, 2023.
- [17] GE Y Q, HUA W Y, MEI K, et al. OpenAGI: when LLM meets domain experts[J]. arXiv: 2304. 04370, 2023.
- [18] ETSI. Zero-touch network and service management (ZSM) (Version 2.1.1)[S]. 2023.
- [19] ETSI. Generic autonomic network architecture (GANA) (version 1.1.1)[S]. 2020.
- [20] TMF. Autonomous networks whitepaper (3.0)[R]. 2021.
- [21] 刘贤松, 陈凌, 尹俊. 通信行业语言大模型技术和应用研究[J]. 邮电设计技术, 2023(7): 14-19.
LIU X S, CHEN L, YIN J. Research on large language model technology and application in communication industry[J]. Designing Techniques of Posts and Telecommunications, 2023(7): 14-19.
- [22] 毛志伟, 邢向晖, 孙广生. 浅谈 AIGC 在通信设计领域的应用[J]. 邮电设计技术, 2023(7): 25-30.
MAO Z W, XING X H, SUN G S. Discussion on application of AIGC in communication design field[J]. Designing Techniques of Posts and Telecommunications, 2023(7): 25-30.
- [23] 李涛, 王春佳, 李姗姗. 电信网络智能化方案研究[J]. 电信科学, 2023, 39(3): 162-172.
LI T, WANG C J, LI S S. Research on intelligent scheme of telecommunication network[J]. Telecommunications Science, 2023, 39(3): 162-172.

[作者简介]



廉霄兴 (1990-), 男, 亚信科技 (中国) 有限公司数据智能规划高级专家, 主要研究方向为人工智能算法、大数据及数字孪生技术。



宋勇 (1989-), 男, 亚信科技 (中国) 有限公司算法工程师, 主要研究方向为行业大模型开发应用、自然语言处理开发。



朱军 (1972-), 男, 博士, 亚信科技 (中国) 有限公司高级产品总监、首席架构师, 主要研究方向为数据智能产品规划、大数据平台架构设计、企业数智化转型。



王淑玲 (1988-), 女, 博士, 亚信科技 (中国) 有限公司研发中心规划部规划总监, 主要研究方向为网络通信、云网融合。



叶晓舟 (1980-), 男, 博士, 亚信科技 (中国) 有限公司通信人工智能实验室资深总监、首席科学家, 主要研究方向为通信网络与人工智能。



欧阳晔 (1981-), 男, 博士, 亚信科技 (中国) 有限公司首席技术官、高级副总裁, 主要研究方向为移动通信、人工智能、数据科学、科技研发创新与管理。