

面向体系化AI运营的实时能力交易机制研究与应用

赵鹏¹, 蔡海涛¹, 李金翼¹, 唐国华¹, 史湘荧², 袁向阳¹, 王斌¹, 邓超¹, 冯俊兰¹

1. 中移九天人工智能科技(北京)有限公司(九天人工智能研究院), 北京 100033;

2. 北京邮电大学电子工程学院, 北京 100083

摘要

在开放复杂的算网环境下, AI与算力网络一体化服务需求增长, 模型与资源的高效匹配成为亟待突破的瓶颈。为此, 针对复杂任务场景下的体系化AI操作系统(HAI OS), 设计了适配AI能力与算网资源特性的交易机制, 通过对算力、模型和数据的动态感知与统一编排, 在供需两端实现高效实时匹配, 按需将AI能力分发至相应计算与传输节点; 通过AI能力交易模块的研发应用, 推动HAI OS实现去中心化的AI能力智能调度与运营, 为复杂任务和多样化应用场景提供强有力的支撑。

关键词

体系化人工智能; 体系化人工智能操作系统; 实时能力交易; 智能体

中图分类号: TP391

文献标志码: A

doi:10.11959/j.issn.2096-0271.2026042

Research and application of real-time capability exchange mechanisms for holistic AI operations

Zhao Peng¹, Cai Haitao¹, Li Jinyi¹, Tang Guohua¹, Shi Xiangying², Yuan Xiangyang¹, Wang Bin¹, Deng Chao¹, Feng Junlan¹

1. China Mobile Jiutian Artificial Intelligence Technology (Beijing) Co., Ltd., Beijing 100033, China

2. School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100083, China

Abstract

In the open, complex computing-network environment, the demand for integrated AI and computing-network services has been increasing, and the efficient matching of models and resources has become a critical bottleneck. To address this, for the hierarchical AI operating system (HAI OS) in complex task scenarios, a transaction mechanism was designed that adapts to the characteristics of AI capabilities and computing network resources. By dynamically perceiving and orchestrating computing power, models, and data, the mechanism realized efficient real-time matching between supply and demand. It distributed AI capabilities on demand to the corresponding computing and transmission nodes to accomplish specific tasks. Through the development and application of the AI Exchange module, HAI OS achieved decentralized intelligent scheduling and operation of AI capabilities, providing strong support for complex tasks and diverse application scenarios.

Key words

holistic artificial intelligence, holistic artificial intelligence operation system, real-time capability exchange, agent

0 引言

近年来,人工智能(artificial intelligence, AI)技术迅猛发展并成为科技变革的核心动力。《“十四五”国家信息化规划》《“十四五”数字经济发展规划》《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》等文件鼓励构建多层次、多场景的人工智能赋能生态。在政策牵引、技术演进与市场需求的三重驱动下,中国移动将人工智能纳入公司发展战略,打造体系化人工智能操作系统(holistic artificial intelligence operating system, HAI OS),其具备任务分解、原子化模型探索与调度等一系列核心闭环能力^[1-4],可实现算力、网络、AI能力的统一调度和开放服务,促进人工智能像通信服务一样普惠泛在^[5]。

本文面向AI服务、算网资源融合场景,研究以大模型技术为核心的AI与算网多要素能力融合交易方案,并进行实践验证,以实现AI服务需求与AI能力、算网资源的优化匹配^[6-7]。主要工作包括:基于HAI OS框架,提出AIX(AI Exchange)概念与体系定位;设计AIX交易机理与实现方案,支持多方主体间多要素能力的安全、公平、高效流转;结合AI、算力、网络一体化运营实践验证所提出交易机制的可行性。

1 相关研究

1.1 体系化人工智能操作系统

体系化人工智能操作系统定位为体系

化人工智能的智能调度和运营管理的核心系统,在动态环境中对AI能力和算网资源进行高效调度,持续为用户提供即插即使用的AI服务^[8]。现有工作主要围绕体系化AI框架下异构资源与能力的服务规划、检索与串联。本文在此基础上引入自动化竞价机制,将资源调度扩展为去中心化能力交易,实现多要素能力的择优匹配与动态定价,提升能力供给质量。

1.2 HuggingGPT

HuggingGPT^[9]是以LLM(ChatGPT)为控制器,用专家模型(Hugging Face)执行任务的AI Agent系统,将自然语言作为接口调度各类专家模型完成复杂的AI任务^[10]。相关研究工作重点在于规划LLM与外部模型协作的AI领域方案。本文借鉴相关思路提出AI能力协作框架,区别于仅依据模型下载量的简单排序机制选择模型^[11],本文在能力匹配阶段提出更高效精准的动态竞价机制,综合考虑AI模型性能、资源成本及用户需求适配性,显著提升多模型协同的调度效率与匹配精度^[12]。

1.3 Ad Exchange

广告交易(advertising exchange, Ad Exchange)平台是连接广告交易买方和卖方的在线交易市场,通过实时竞价实现广告资源交易^[13]。广告交易的主要流程通常涉及供给方平台对广告位及用户数据进行实时推送、Ad Exchange发起竞价请求、需求方平台依据广告主出价策略参与竞价等。市场上,腾讯的Tencent ADX、阿里妈妈的Tanx等均形成了高度协作的

广告交易闭环。相关研究重点在于实时竞价与自动化交易，本文借鉴其思路提出针对AI能力交易的AI Exchange架构，将模型性能、算力资源成本等多维因素纳入统一竞价空间，设计能力竞价、竞优交易路径，构建开放透明、公平竞争的AI能力交易环境^[14]。

2 AI Exchange定义和意义

2.1 定义

AI Exchange作为体系化智能调度和控制系统的核心平台，对算力、模型、数据等AI资源进行动态感知与统一编排，在AI能力的供需两端之间建立高效的实时连接与匹配，按需调度AI能力至相应的计算与传输资源节点以完成相应的任务，实现去中心化的体系化智能调度和运营^[15]。完整过程如图1所示。

上述过程涉及的角色包括：AI服务需求方，指提出并使用智能化服务需求的终端主体，涵盖B、C端用户（企业客户和个人客户）；AI服务提供方，作为核心调度中枢负责对接需求方与能力提供方，实现AI能力的精准匹配、完成AI服务的编排调度与执行^[16]；HAI OS，作为系统的智能调度与管控操作系统，具备能力管理、资源调度、服务编排、交易撮合与策略推理等核心功能，支撑整个AI服务生态的系统级协同与治理；能力汇聚平台，面向AI能力提供方提供统一注册、管理、竞价与交易支撑的服务门户^[17]，具备多租户隔离、标准协议接入、能力上下架管理等核心功能；能力提供方，指提供AI模型、算法服务的第三方主体，通过标准化接口向AI Exchange注册能力元数据，如模型描述、性能指标等，并参与动态竞价与交易过程，实现能力的市场化运营。

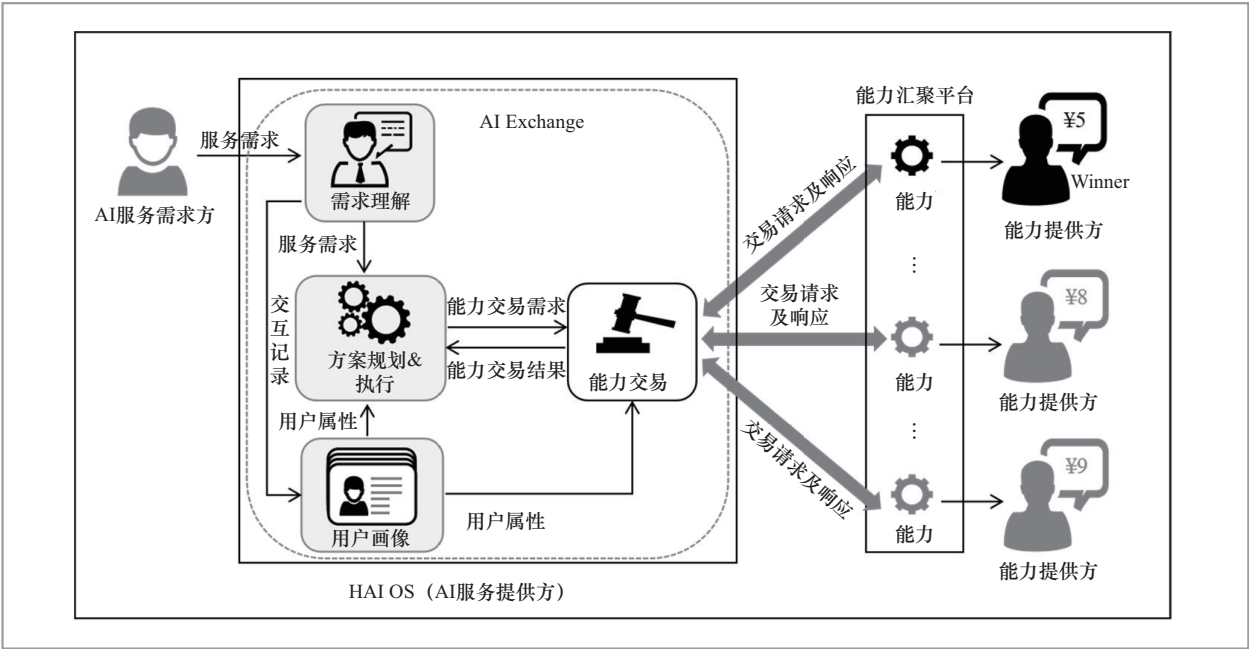


图1 完整过程示意图

2.2 意义

AI Exchange 构建去中心化的能力交易生态，核心目标是通过平台化机制高效匹配用户需求，既支持单要素能力的订购，也满足复杂解决方案的联合订购。平台通过解耦需求方、能力方与算力网络资源，促进多要素独立协同并规范服务体系；建立开放透明的交易机制与统一标准，营造公平竞争环境^[18]；完善交易与收益分配机制，支持多维度动态定价，提升能力供给方的参与度；持续引入多领域 AI 能力与资源，依托智能匹配与动态竞价优化供给广度与质量，构建体系化 AI 服务供给体系，推动形成高竞争力的智能生态。

3 AI Exchange 机理

3.1 能力交易模式

在 AI 服务提供方与能力提供方的实时能力交易中，根据竞争维度的不同，存在竞价和竞优两种方案：竞价方案中标的能

力明确且统一，各参与方只竞争价格；竞优方案需要由能力提供方自己推选能力参与竞争，并最终由 AI 服务提供方综合择优选取^[19-20]。

竞价方案还包括一次出价判定交易结果的单次能力竞价方案和参考竞拍过程，以及参与方多次出价竞争的多轮竞价方案，前者简单、高效，后者则能获得更有竞争力的成交价格。

3.1.1 单次能力竞价

单次能力竞价方案的特征首先在于只竞价：各能力提供方之间针对明确的标的（某个能力）只进行价格上的竞争，参与竞价的标的能力是由 AI 服务提供方根据用户需求匹配选定而非由能力提供方匹配选择；各能力提供方只需进行一次性报价，AI 服务提供方据此进行价格比对。在 AI 服务运营业务场景中，出价低的能力得到提供服务的机会。实时能力交易的单次能力竞价方案如图 2 所示。

作为整体方案的前置条件，首先各能

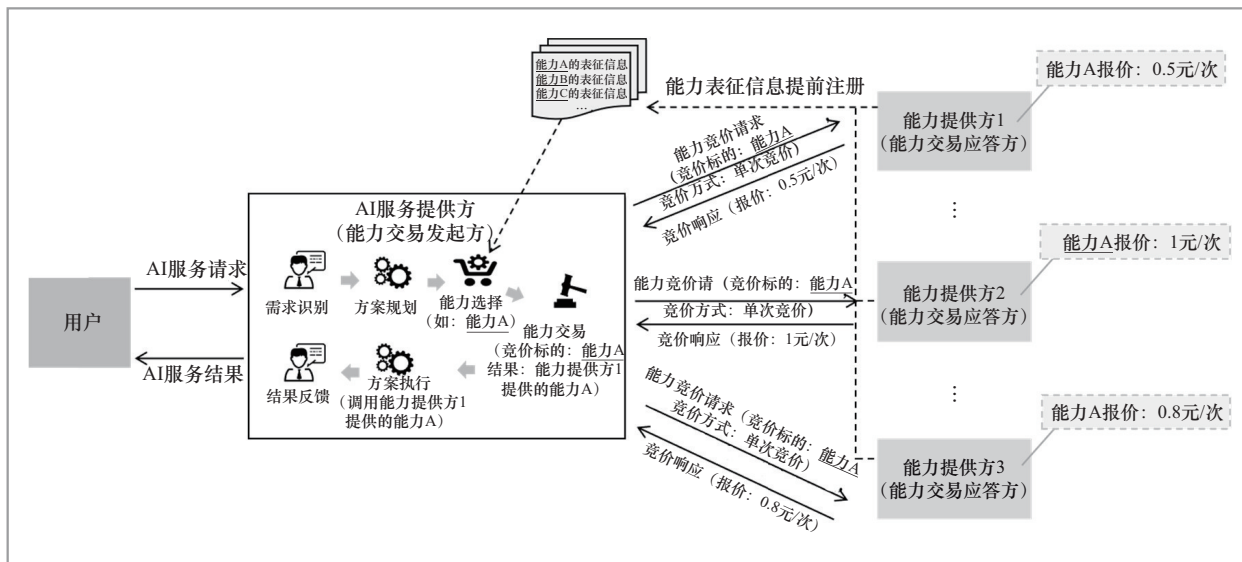


图2 实时能力交易的单次能力竞价方案示意图

力提供方需要提前将自己提供的所有能力信息（包括能力基本信息、功能、性能等表征信息）上报给 AI 服务提供方的服务平台（如图 2 所示，能力提供方 1、2、3 将能力 A、B、C 的信息提前上报给 AI 服务提供方），以便后续服务提供方进行能力匹配和选择。

当用户发起 AI 服务请求时，AI 服务提供方通过大模型理解并拆解用户需求，根据用户需求规划生成解决方案，拆解的任务中包含需要哪些能力及其组合来实现服务。这里可能包含多种能力需求，AI 服务提供方需要逐一对这些能力需求发起能力交易。

能力交易的过程是，AI 服务提供方依据能力需求通过大模型在已有的能力表征信息中选择满足方案需求的能力（如图 2 所示，AI 服务提供方选择了能力 A），然后向能力所属提供方发起竞价请求。请求中已经指明了所需的能力和单次能力竞价的竞价方式。能力提供方需要针对指定的能力进行报价，并反馈 AI 服务提供方（能力提供方 1、2、3 针对能力 A 分别报价 0.5 元/次、1 元/次、0.8 元/次调用）。AI 服务提供方选择出价最低的某个能力提供方提供的能力（最终成交为报价 0.5 元/次调用的能力 A），完成能力的部署调用，并向用户提供 AI 服务。

假设能力提供方的集合为 $N=\{1,2,\dots,n\}$ ，每个能力提供方的出价向量为 $b=(b_1,b_2,\dots,b_n)$ ，其中 b_i 表示能力提供方 i 的出价，则交易成功的能力提供方为：

$$N_{\text{成交}} = \arg \min_{i \in N} b_i \quad (1)$$

交易成交价为：

$$P_{\text{成交}} = \min_{i \in N} b_i \quad (2)$$

单次能力竞价方案的优点是通过能力提供方的一次出价即可得出交易结果，方

案简单，流程和接口简洁，对于能力提供方来说，实现逻辑也更加简单。但是单次能力竞价的最终交易价格可能不是最具竞争力的价格。

3.1.2 多轮能力竞价

多轮能力竞价采用多轮出价的方式进行交易价格撮合，能力提供方可以按照自身定义的出价策略进行多次出价^[21]，而 AI 服务提供方则扮演“竞拍”主持角色，组织整个多轮次出价的竞价过程并最终决策。实时能力交易的多轮竞价方案如图 3 所示。

当 AI 服务提供方选择出满足方案需求的能力后（如图 3 所示，AI 服务提供方选择了能力 A），向能力所属能力提供方发起竞价请求。请求中指明竞价方式为多轮竞拍。能力提供方针对指定的能力基于其他能力提供方的报价和自身的策略进行多次报价，并反馈给 AI 服务提供方，AI 服务提供方负责将当前的最低价通知给所有能力提供方，等待能力提供方报出低于当前报价的价格，如果没有新的出价，则以当前最新的报价结束交易（如图 3 所示，第一轮能力提交方 1、2 分别报价 0.5 元/次、0.8 元/次调用，AI 服务提供方通知能力提供方当前最低价是 0.5 元/次调用并等待新的出价，第二轮能力提供方 2 报出新的出价 0.4 元/次调用，能力提供方 1 没有再出价，因此能力提供方 2 以 0.4 元/次调用的出价赢得交易）。AI 服务提供方根据交易结果完成能力的部署调用，并最终向用户提供 AI 服务。

假设能力提供方的集合为 $N=\{1,2,\dots,n\}$ ，交易初始价格为 P_0 ，第 t 轮的报价为能力提供方提交 $b_i^{(t)}$ ，满足 $b_i^{(t)} < b_i^{(t-1)}$ （递减拍卖）， T 为拍卖终止轮次（即仅剩一名参与者继续出价时对应的最终轮次）。终止条件与成交规则为当仅剩一位参与者继续出价

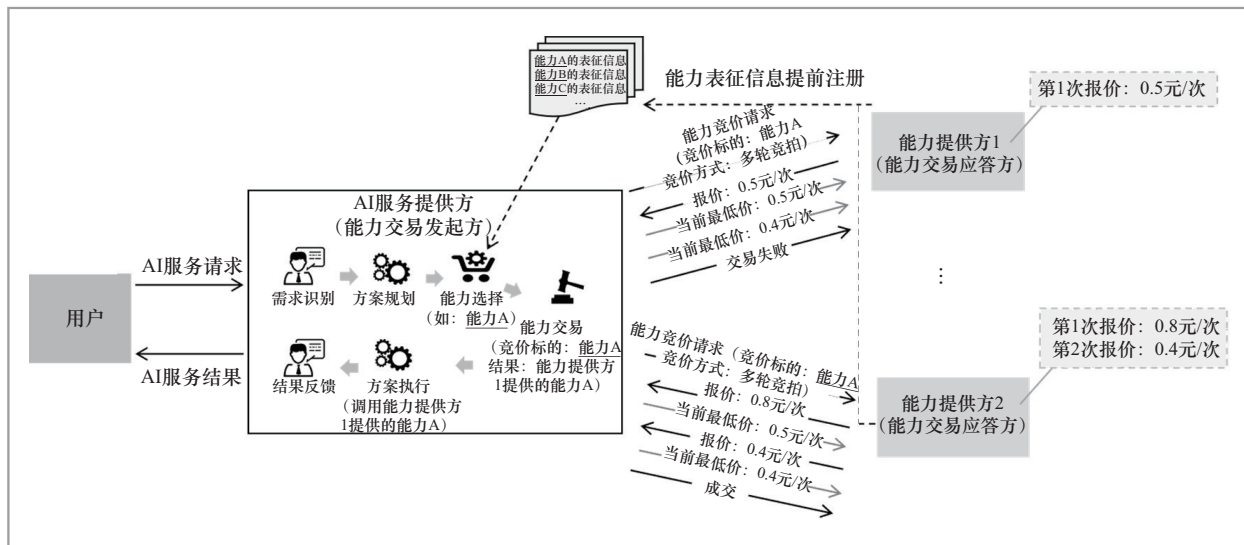


图3 实时能力交易的多轮竞价方案示意图

时，其以当前轮次最高报价成交：

$$\exists! i \in N \text{ 满足 } b_i^{(T)} \leq P_T \Rightarrow P_{\text{成交}} = b_i^{(T)} \quad (3)$$

多轮能力竞价方案的优点是允许能力提供方通过多次出价探索最优报价，动态优化报价，通过竞拍的方式让交易价格更具竞争力。但交易过程相对复杂，交易过程的时间长，特别是在一次 AI 服务涉及的能力较多的场景，对 AI 服务的响应时间影响较大。

3.1.3 能力竞优

为了在更大范围内且更加灵活地匹配能力，能力竞优方案将部分能力选择的主动权让给能力提供方，由能力提供方自主选择能力进行交易并报价。AI 服务提供方不再只根据价格竞选能力，而是综合推选能力的价格、性能、质量等多种因素动态、择优选取。实时能力交易的能力竞优方案如图 4 所示。

作为整体方案的前置条件，首先各能力提供方需要提前向 AI 服务提供方注册能

力提供方信息，包括能力提供方基本信息、所支持的能力范围信息等，但无须提前提供或注册具体的能力表征信息。

当一个用户发起 AI 服务请求时，AI 服务提供方规划方案并拆解能力需求；然后面向支持范围匹配需求的所有能力提供方广播能力交易需求，需求中包含详细的能力需求描述、AI 服务信息、服务对象（用户）信息等，用于能力提供方进行能力选择。能力提供方根据需求信息，结合自身策略，匹配自有能力并进行报价（如图 4 所示，能力提供方 1 推选能力 A，处理时延为 10 ms，报价 0.8 元/次；能力提供方 2 推选能力 B，处理时延为 20 ms，报价 0.8 元/次；能力提供方 3 推选能力 C，处理时延为 50 ms，报价 0.5 元/次）。AI 服务提供方根据交易策略最终选择使用某个能力提供方提供的能力（如图 4 所示，能力提供方综合价格和性能因素，选择能力提供方 1 提供的能力 A 来提供服务），完成能力的部署调用，并最终实现用户的 AI 服务。

能力竞优方案的优点是能力选择是逐需求的动态匹配，灵活性更高，由于参考

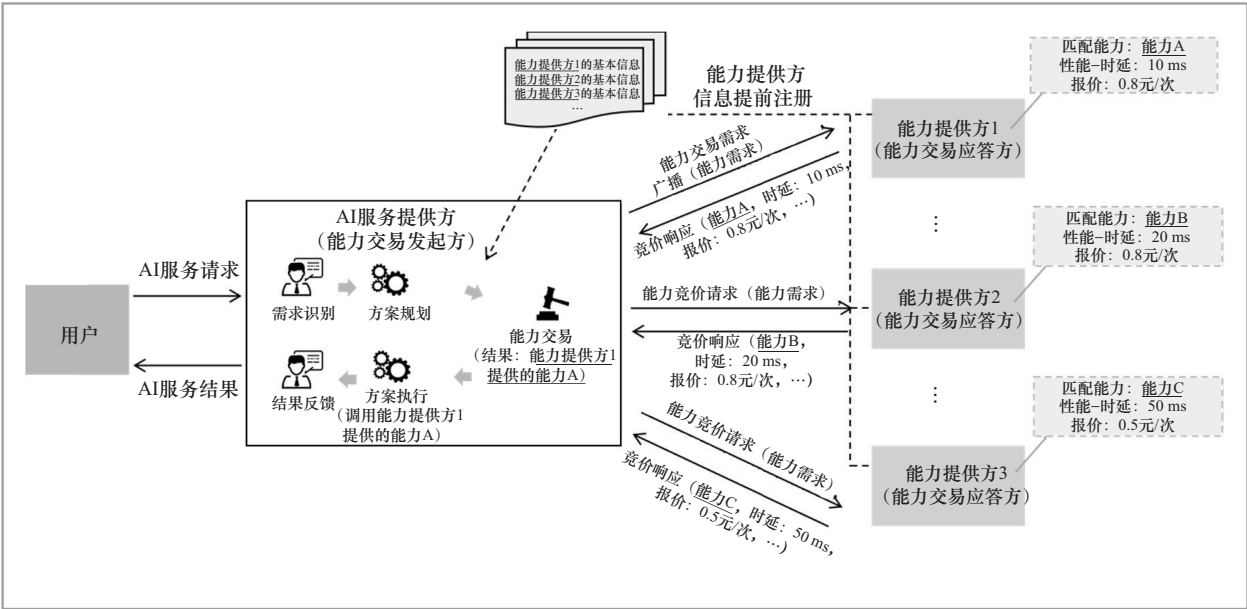


图4 实时能力交易的能力竞优方案示意图

因素更丰富，能够更精确匹配用户需求。但是其需要所有能力提供方参与能力选择的过程，还要保持 AI 服务提供方与能力提供方的策略一致，对能力提供方提出了更高的要求。

3.2 能力交易策略

能力交易中能力选择和交易决策的依据来自于交易策略，基于为用户提供更高质量 AI 服务的目的，能力交易策略的核心是择优而非单纯的价格比较。因此，在能力交易过程中需要综合多维度的因素，支持更加灵活的交易策略，在静态策略决策的基础上，支持大模型、智能体等智能化手段增强交易决策综合分析能力，提升策略实施的准确性。

3.2.1 应用场景

在能力竞价方案中，能力交易结果判

定的依据只有价格，但交易标的物的选择需结合交易策略综合选取。交易策略的应用是在正式交易前的能力选择过程中，由 AI 服务提供方负责实施，能力提供方只参与价格竞争，不参与能力选择，不涉及交易策略的实施。在能力竞选优方案中，首先由能力提供方根据需求并结合自身交易策略推选参与竞优的能力，AI 服务提供方依据能力交易策略决策最终的采用方。为减少参与各方之间的耦合和依赖，AI 服务提供方和能力提供方都可设定自身交易策略，保持各自策略独立，互不影响。

3.2.2 策略影响因素

能力交易策略需要综合考虑多方面因素，除了能力本身的价格、质量和性能等因素，还要考虑能力提供方本身的服务能力，此外，还需要结合当前用户的定位或等级来制定针对性的交易策略，从而更准确地匹配用户需求，更高效地提供高质量

AI 服务。
常用的交易策略因素见表1。

3.2.3 策略实施

在能力竞价方案的能力选择阶段或者能力竞优方案的交易结果裁定阶段，涉及交易策略的实施，AI 能力提供方需要实时地根据各决策影响因素的情况，综合计算并给出能力选择或交易裁定的结果^[22]。

(1) 静态策略算法

一般业务场景下，交易策略影响因素相对简单且固定，可以采用静态策略算法进行计算。但不同 AI 服务场景或者不同类型的能 力，其决策影响因素存在差异，可以通过预置针对不同场景的参数配置（如各影响因素的权重）模板来解决。

假设候选能力集合为 $C=\{c_1,c_2,\cdots,c_n\}$ ， c_i 表示第 i 个能力；用户属性（如高 ARPU、价格敏感、高黏性等）为 u ，用户属性对应的权重向量为 $W_u=[w_p,w_q,w_{perf},w_s]$ ，分别表示用户属性 u 在报价、质量、性能和供应商资质等策略

因素的权重，满足 $\sum w=1$ 。
 b_i 为能力 c_i 的出价，因为出价越低得分越高，所以需逆向指标正向化，并归一化到 $[0,1]$ 区间， $b_{\max}$ 、 $b_{\min}$ 为最高出价和最低出价，则能力 c_i 在价格维度的打分为：

$$B_i=\frac{b_{\max}-b_i}{b_{\max}-b_{\min}} \tag{4}$$

Q_i 、 $Perf_i$ 和 S_i 分别表示能力 c_i 在质量、性能和供应商资质 3 个方面的打分。其计算方法均采用加权综合的方式，以质量为例：

$$Q_i=\alpha_n\cdot\frac{N_i}{N_{\max}}+\alpha_f\cdot\frac{F_i}{F_{\max}}+\alpha_r\cdot\frac{R_i}{R_{\max}} \tag{5}$$

N_i 表示能力 c_i 历史调用次数， F_i 表示能力 c_i 历史调用频次（次/单位时间）， R_i 表示用户对能力 c_i 的评分。 $[\alpha_n,\alpha_f,\alpha_r]$ 是能力质量各指标对应的权重向量，满足 $\sum\alpha=1$ 。

最后，通过所有策略因素的加权综合计算，获得的能力为：

表1 常用策略影响因素

策略因素	说明	评估指标	策略倾向
能力价格	能力提供方在能力交易过程中所提供的能力报价	直接价格比较	越低越好
能力质量	能力的质量,在没有实际的横向对比测试数据的情况下,可以参考能力的历史调用数据和评分	历史调用次数、历史调用频次、用户评分等	越高越好
能力性能	不同能力的性能表征不同,而且需要结合能力本身的功能需求来对能力的全部或部分性能进行评估	响应时延、资源效率、单位时间调用次数、处理量等	越高越好
能力提供方资质	能力提供方服务能力和质量的考虑因素,可以从与该供应商的历史合作数据、该供应商本身的服务数据等方面进行评估	提供商能力供应规模、历史总调用次数、历史调用频次、用户打分等	越高越好
用户属性	用户的属性也影响能力交易的策略,如高 ARPU 用户偏向服务体验更好的策略,但价格不一定最便宜;一般用户在满足需求的情况下对价格可能更加敏感。又如,向黏性高的用户提供更好体验的服务,而向黏性低的用户提供成本更低的服务等	用户 ARPU、VIP/非 VIP、用户等级、用户黏性等	根据属性不同而不同
保护因素	防止恶意出价的保护策略等	无	无

$$C_{成交} = \arg \max_{c_i \in C} (w_p \cdot P_i + w_q \cdot Q_i + w_{perf} \cdot Perf_i + w_s \cdot S_i)$$

(6)

从以上计算可见，除策略影响因素的计算外，还需根据不同业务场景和影响因素配置权重参数。

(2) 智能决策

静态策略的计算准确性依赖于策略影响因素的多少和权重等预置参数设定是否合理。面向复杂的场景，静态策略计算会有较大局限，扩展性也不高。面向能力交易的复杂场景，可以利用人工智能大模型或 Agent 的综合分析和推理能力来支撑能力交易的决策。

3.3 个性化服务的能力交易

在体系化 AI 交易机制中，除了在供给侧（面向能力提供方）通过实时竞价与竞优提升能力供给质量，还需在服务侧（面向用户）增强定制化与个性化供给能力，以提升整体服务体验^[23]。

用户个性化信息主要包括用户优先级划分、偏好标签、历史调用数据及满意度反馈等^[24]。该信息由 AI 服务提供方在服务过程中采集与分析，通过统计分析 with 模式识别方法，挖掘隐性需求与偏好特征，以支持交易策略的优化与能力选择的智能化。在提供个性化服务的同时，系统遵循最小必要原则，对用户数据进行脱敏与权限分级管理，结合差分隐私与联邦学习等技术强化数据安全，保护用户隐私，确保用户身份信息不被暴露。

3.4 交易可视化

为提升能力交易过程的透明度与可信度，AI Exchange 设计了可视化模块，实现交易过程的实时展示与历史追溯。用户

与能力提供方均可查看交易需求、竞价方式、出价结果等关键信息，从而确保交易的公开、公平与可验证。

在展示过程中，系统严格执行隐私保护机制。具体包括：在 AI 服务提供方的交易过程直接 UI 展示中隐藏参与方的身份标识（identity document，ID）等信息，同时每个能力提供方通过接口从 AI 服务提供方获取的交易信息中也只能看到自己的 ID 信息，仅能访问自身交易记录，避免跨方数据泄露，其他能力提供方的 ID 信息需被隐藏。

可视化模块同时记录完整的交易日志，支持后续的服务评估与策略优化，为体系化 AI 运营提供透明化的过程支撑。

4 AI Exchange 实现方案

4.1 实现框架

AI Exchange 基于体系化智能调度理念，整体设计实现了从用户需求解析、任务拆解、能力竞价交易到执行优化的完整闭环^[25]。整体架构如图 5 所示，用户通过服务入口提出业务请求；运营管理模块负责对用户请求进行意图拆解和需求分析，生成标准化、结构化的服务需求信息；交易市场模块基于用户需求与能力特征进行能力筛选与竞拍控制；能力汇聚平台作为供给侧管理的核心，支持 AI 能力的上架、测试、变更与交易；编排调度中心通过智能中枢统一管理算力与网络资源，完成任务与资源的自动规划与动态调度。HAI OS 作为系统的大脑，整合大模型的推理能力，实现任务规划、能力匹配、资源调度与优化执行的智能决策与实现，保障整个交易和执行过程的智能化与高效性。

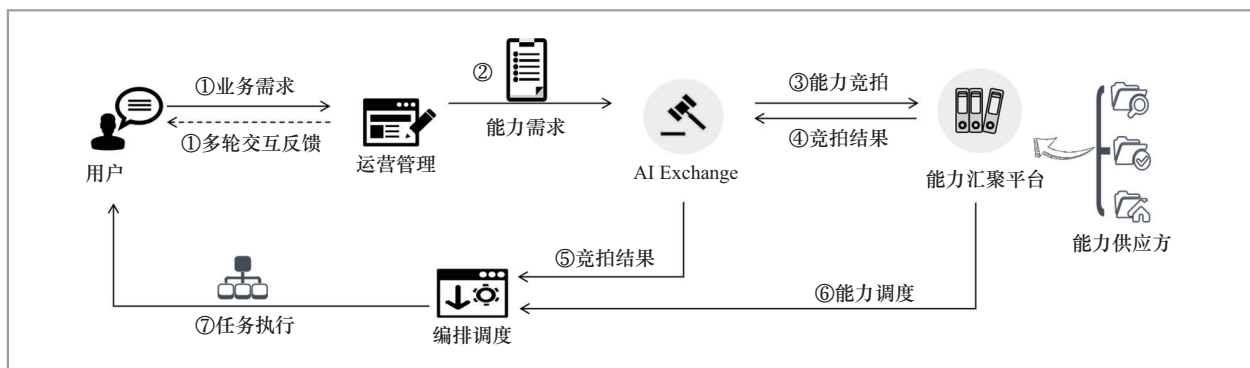


图5 AI Exchange系统架构

在整体流程层面，以用户需求为驱动，从服务入口到最终结果反馈，设计以下流程。

- ①用户需求提出：用户在运营入口通过自然语言交互的方式发起服务请求。
- ②意图解析：运营管理模块与用户多轮交互，理解用户意图并进行需求分析。
- ③任务规划：由HAI OS智能中枢根据意图分析的结果，将需求拆解为多个子任务^[26]，明确任务目标、资源要求与能力匹配标准。
- ④能力选取：系统在能力汇聚平台中检索匹配符合要求的AI能力，将规划好的任务与能力列表中的AI能力进行匹配，确定每个任务需要使用哪些AI能力。
- ⑤能力交易：交易市场模块发起能力竞价，对所需的AI能力进行竞拍。采用能力竞价或竞优两种交易方案，支持单轮或多轮动态竞价，最终返回竞价结果。
- ⑥端到端优化：竞价完成后，系统依据竞价结果选定能力，对选定的执行方案进行端到端的优化，包括对能力组合、参数调优、资源协同规划的可执行方案进行再训练与优化。
- ⑦算网调度：在编排调度阶段，HAI OS调用相应的AI能力，结合算力与网络规划，合理分配和调度所需资源，确保执行方案的性能与效率，提升任务执行的整体性能。
- ⑧模拟验证和反馈：在执行方案正式实施前，通过模拟验证机

制对执行方案进行预评估，确保方案的可行性和有效性，及时发现潜在问题。根据模拟验证的结果对方案进行优化，提升方案执行效果。

- ⑨任务执行：任务执行模块依据优化后的方案调度各项AI能力完成任务。
- ⑩总结、输出：任务结果交由总结输出模块处理并反馈给用户，形成闭环式的服务体验。用户可对服务结果进行确认与反馈，进一步完善系统的持续优化能力^[27]。整体流程如图6所示。

4.2 管理流程和接口

4.2.1 能力管理流程与接口

能力管理是AI Exchange平台对供给侧能力的全生命周期管控，包括注册、注销、变更与查询。所有待交易能力需先在HAI OS注册，便于平台识别、匹配和调用。能力管理流程包括以下3类典型场景。

- 能力注册与注销：能力提供方首次接入平台或新增能力时，需要通过注册接口提交能力名称、描述等信息；若能力停用或停止维护，提供方通过注销接口将其标记为失效。

- 能力信息变更上报：能力版本升级、性能优化等信息发生变化时，提供方更新

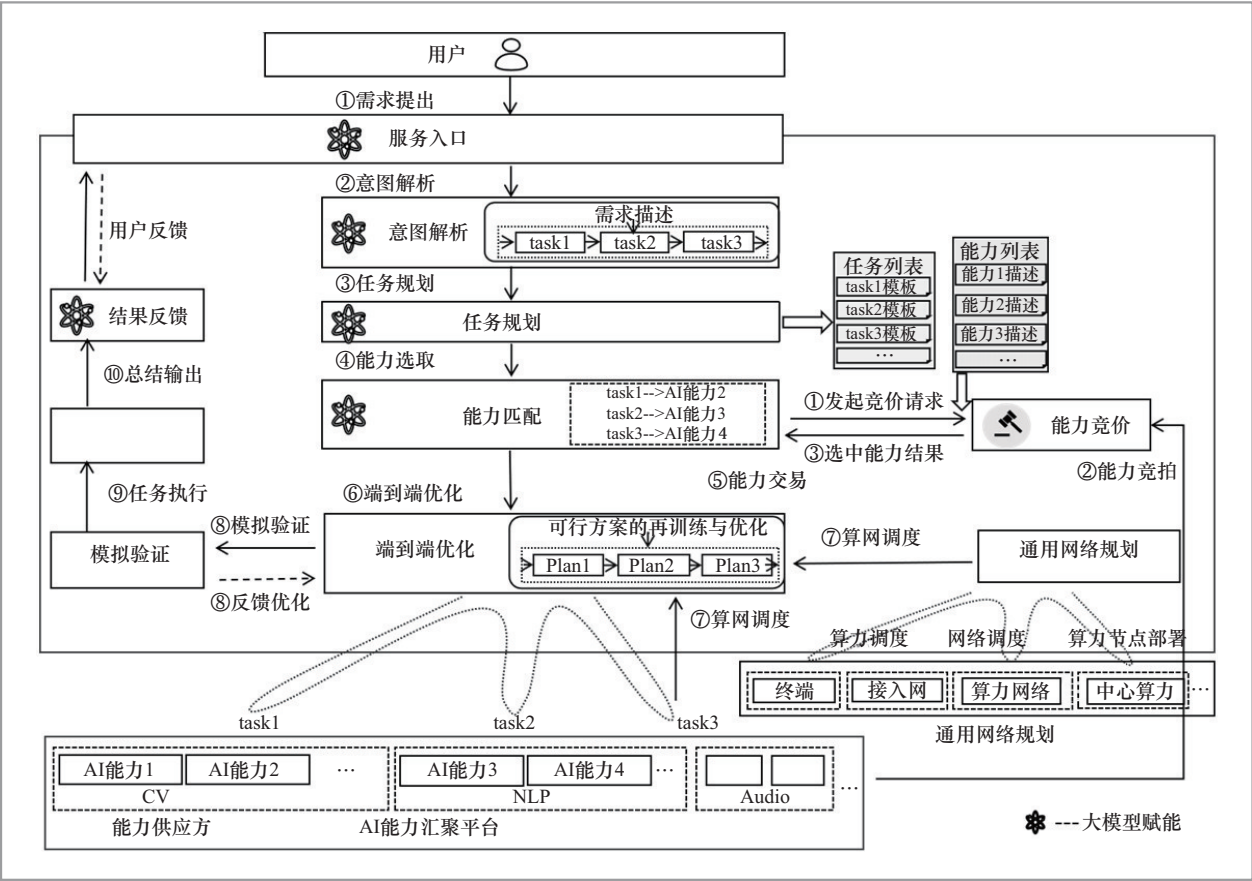


图6 AI Exchange整体流程

接口上报变更内容及能力标识，HAI OS接收请求后及时更新能力数据库并同步新的状态。

- 能力信息查询：HAI OS可向提供方主动发起查询请求，既能一次性获取全部能力的列表信息，也能按能力标识获取指定能力的完整细节。

4.2.2 能力提供方管理流程与接口

能力提供方是AI Exchange供给侧的关键角色。平台通过统一接口对所有能力提供方执行注册、变更、查询和注销操作。能力提供方管理流程包括以下3类典型场景。

- 能力提供方注册与注销：首次接入

时能力提供方提交基本信息、服务范围、资质说明与状态，系统分配唯一提供方ID并完成登记。若提供方暂停业务，需发起注销请求。

- 能力提供方信息变更上报：当能力提供方的基本资料、业务范围或资质状态发生变化时，可主动上报更新内容及ID，HAI OS及时修改数据库并反馈结果。

- 能力提供方信息查询：平台可定期向能力提供方发起查询，获取最新状态、资质及服务能力。

4.2.3 能力交易流程与接口

HAI OS与能力提供方通过统一接口实

时交互，依据是否预先指定候选能力形成“竞价”和“竞优”两种模式。借助统一接口与双模式机制，平台兼顾结构化采购与动态优选需求；通过轮次控制、策略绑定、历史出价与能力表征反馈，提升交易过程的透明度、灵活性和决策精度，为后续调度与闭环管理奠定数据基础。

(1) 能力竞价

当平台已明确目标能力时，会先从注册能力池筛选符合需求的提供方，组成候选集并发起交易请求。提供方可一次性报价，

也可在多轮竞价中根据平台同步的当前最低价调整出价；平台可设定最大轮次或在价格趋稳时自动终止并确定中标方。接口允许配置能力需求、用户信息、预算、交易策略与轮次数，返回交易ID、轮次状态、各方报价及最终成交能力，用于后续部署与可视化。能力竞价接口参数设计见表2。

(2) 能力竞优

若平台仅提供功能需求而不指定能力项，则向多家提供方广播交易请求。各方根据自身策略判断能力池是否满足需求，

表2 能力竞价接口

接口参数		属性	描述
	request_id	String	请求标识ID
	auction_id	String	多轮竞价会话ID(多轮竞价时必须填)
	bid_request_id	String	当前轮次竞价请求ID
	providers_id	String	指定的能力提供方ID列表
ai_requirements	function	String	能力需求
	user_id	String	用户信息
	strategy_id	String	策略ID
previous_bids	provider_id	String	前轮次的出价信息:能力提供方ID(多轮竞价时可选)
	bid_price	Number	前轮次的出价信息:价格(多轮竞价时可选)
	budget	Number	预算
	max_rounds	Number	最大竞价轮数(多轮竞价时可选)
	round_id	Number	当前轮次(多轮竞价时可选)
	transaction_id	String	本次交易ID
	auction_id	String	多轮竞价会话ID(多轮竞价时返回)
	round_id	Number	当前轮次ID(多轮竞价时返回)
	next_round	Boolean	是否有下一轮竞价(多轮竞价时返回)
	end_auction	Boolean	是否结束竞价(多轮竞价时返回)
bid_response	bid_price	Number	当前竞价
	provider_id	String	能力提供方ID
	final_price	Number	最终成交价(交易成功时返回)
	status	String	交易状态(success/failure/pending)
	estimated_latency	Number	预计时延
	win_lose_notice	String	中标/未中标通知

若满足即可上报能力标识、性能表征与报价。平台依据性能、价格、时延等权重综合评分确定中选能力，并向所有参与方通报结果。能力竞优接口参数设计见表3。

4.2.4 可视化流程与接口

HAI OS 构建了贯穿交易全程的可视化信息交互机制，支持交易中的实时状态同步和完成后的历史记录查询，方便参与方随时掌握竞争进展，并进行事后分析与策略改进。实时同步与历史查询协同实现了交易过程的“实时可观、事后可查”^[28]，增强了提供方对交易流程的信任与参与感，并为平台的能力画像、策略优化和智能运营奠定数据基础。

• 实时信息同步接口：平台通过实时同步接口向所有参与方推送最新竞价数据，包括当轮各方报价、报价时间、最高与最

低价、轮次编号及当前交易状态，并附带时间戳供同步与日志记录。实时推送使提供方能够基于市场动态快速调整出价策略，提高中标机会。

• 历史信息查询接口：交易结束后，提供方可调用历史查询接口按交易或能力维度获取完整记录，涵盖交易起止时间、策略 ID、能力需求、各轮竞价详情、最终中选能力与成交价格等。该接口便于对账、策略复盘、服务评估及合作方式改进等。

5 应用实践

5.1 应用案例介绍

图像、文本、语音方面的综合处理是日常生活中常见的且较典型的智能化服务应用场景，本文以此为例进行介绍和分析：

表3 能力竞优接口

接口参数		属性	描述
request_id		String	请求标识 ID
providers_id		String	参与竞优的能力提供方 ID 列表
ai_requirements	function	String	能力需求
	user_id	String	用户信息
	strategy_id	String	策略 ID
budget		Number	预算
transaction_id		String	本次交易 ID
bid_response	bid_price	Number	报价
	provider_id	String	成交能力提供方 ID
	ability_characteristics	String	能力表征信息
	status	String	交易状态(success/failure/pending)
	ability_id	String	被选中的能力 ID
	win_lose_notice	String	中标/未中标通知
selected_ability	final_price	Number	最终成交价(交易成功时返回)
	ability_id	String	成交被选中的能力 ID
	provider_id	String	成交能力提供方 ID

用户想开一家儿童书店，需要 HAI OS 参考用户上传的照片，设计书店标志，给出一段书店的宣传语文案和一段广告宣传音频，并且提供一张书店的室内装修效果图。HAI OS 首先需要分析用户上传的图片，了解书店的装修风格和色彩搭配，然后基于分析结果生成标志、宣传文案、音频和室内装饰图。本案例涉及图生文、文生图、LLM 等多种 AI 能力，HAI OS 通过能力交易完成能力的最终选取和调用。

5.2 技术实现介绍

因为实时能力交易参考并部分模拟真实环境交易市场竞拍的过程，系统在实现上采用基于多智能体的协同架构^[29]，由负责任务编排的协调调度 Agent 与负责策略计算的规划 Agent 分工协作，以解耦调度控制与竞价决策^[30]。如图 7 所示，AIX 在内部定义统一的能力交易消息协议 AIXM (AIX message)，规划与调度 Agent 通过“能力交易任务请求/响应”等进行消息交互，外部模块通过“交易请求/响应”与调度 Agent 对接。当 HAI OS 上层模块生成候选能力集合后，通过 AIXM 向规划 Agent 发送能力交易需求；规划 Agent 生成交易任务并下发调度 Agent，调度 Agent 向各能力方发起交易请求、汇总结果并回传上层模块用于能力部署。异常处理方面，AIXM 在协议层定义统一的状态

码和错误码，将超时未应答、报文格式错误或超预算等请求标记为无效并记录日志；若中标能力在调用或部署阶段失败，则触发降级策略，回退至备选能力或固定能力调度，以保障服务连续性。

5.3 性能评估与结果分析

HAI OS 原型系统实验平台由 6 台虚拟机组成，配置 16 核 CPU 与 64 GB 内存；集群配备 32 块华为昇腾 910B GPU；操作系统采用 BigCloud Enterprise Linux For Euler 21.10 LTS；集群管理与调度由 Kubernetes v1.18.8 负责，实现算力节点的弹性管理与服务自动编排；模型层面，HAI OS 原型系统采用中国移动九天人工智能大模型 JIUTIAN-75B 32K。在上述环境下，本文构造了 50 组多模态 AI 实验任务，按复杂度（以子任务数量、涉及模态种类作为主要衡量标准）分为 3 类，具体构成见表 4。任务请求选取自 HAI OS 原型系统能力目录中的典型服务模板，覆盖图像、文本、语音、图文联动等场景。用户侧模拟约 30 个虚拟用户，以随机方式发起上述任务，并基于交互反馈统计用户满意度等体验指标。

整体统计指标包括平均响应时延、平均任务完成时间、成本系数、GPU 资源利用率、任务执行成功率、用户满意度评分。结果见表 5。其中，成交系数以无 AIX 模

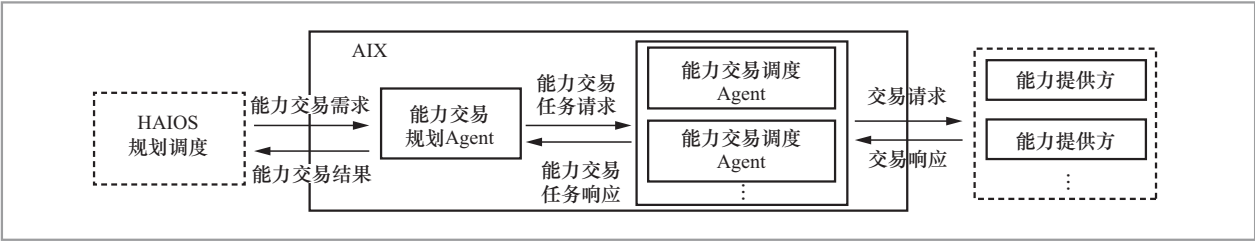


图7 基于AIXM的能力交易消息交互

表4 实验任务按复杂度的构成情况

任务类型	任务数量	子任务数量	涉及模态	典型场景/约束示例
简单任务	20	1~2	单一模态或双模态(如图像+文本)	图像分析+文案生成等基础服务
中等任务	20	3~4	至少两种模态,包含跨模态子任务	图像理解+文本生成+语音合成等组合任务
复杂任务	10	≥4	多模态、多轮交互	在上述组合基础上叠加多轮交互或严格时延约束

表5 不同交易模式运行特征对比

交易模式	平均响应时延/ms	平均任务完成时间/s	成本系数	GPU资源利用率	任务执行成功率	用户满意度评分(1~5)	特征分析
无AIX模式	178	12.3	1	31%	85.2%	3.7	传统模式,成功率最低
单轮竞价	180 ↑1.1%	12.8 ↑0.5	0.905 ↓9.5%	38% ↑7%	86% ↑0.8%	4.0	响应快,价格竞争力一般
多轮竞价	235 ↑32%	14.1 ↑1.8	0.872 ↓12.8%	45% ↑14%	87.2% ↑2%	4.1	多轮报价,成本优化但耗时略长
能力竞优	210 ↑18%	13.5 ↑1.2	0.95 ↓5%	47% ↑16%	89.8% ↑4.6%	4.3	综合择优匹配,成功率高

式在相同任务集合下的平均单次任务成本为1进行归一化。记无AIX模式的平均单次任务成本为 C_{base} ，其他模式为 C_i ，则成本系数统一定义为 $R_i=C_i/C_{base}$ 。从表5看出，相较基线无AIX模式，引入AI Exchange后系统在任务成功率、资源利用率和用户满意度等方面整体提升，多轮竞价模式在成本系数上最具优势，能力竞优模式在多项指标上表现均衡，而单轮竞价模式在响应时延方面保持优势，验证了AI Exchange在成本、性能与体验之间提供可调节折中的能力。

能协同的实现方案，涵盖实现框架、管理流程与接口；通过具体应用案例展示了AI Exchange在实际场景中的落地效果与应用价值，验证了方案的可行性与创新性，为后续规模化推广奠定了基础。笔者团队在体系化AI Exchange技术研究与系统研发方面已取得阶段性成果，但在工程实现、理论突破与算法创新上仍面临诸多挑战。未来将持续迭代升级体系化AI Exchange的运行模式与架构设计，拓展更丰富的应用场景。

参考文献：

[1] Prasad A, Koller A, Hartmann M, et al. ADaPT: as-needed decomposition and planning with language models[C]// Findings of the Association for Compu -

6 结束语

本文深入论述了AI Exchange的定义、发展目标与基本机理，提出面向多要素智

- tational Linguistics: NAACL 2024. Stroudsburg: Association for Computational Linguistics, 2024: 1467–1483.
- [2] Muqeeth M, Liu H, Liu Y, et al. Learning to route among specialized experts for zero-shot generalization[PP]. arXiv preprint, 2024. arXiv:2402.05859.
- [3] Su J, Xia Y, Lan Q, et al. Difficulty-aware agent orchestration in LLM-powered workflows[PP]. arXiv preprint, 2025. arXiv:2509.11079.
- [4] 冯俊兰. 体系化人工智能[J]. 北京邮电大学学报, 2024, 47(4): 1–10.
- Feng J L. Holistic artificial intelligence[J]. Journal of Beijing University of Posts and Telecommunications, 2024, 47(4): 1–10.
- [5] Larsen B, Li C, Teeuwen S, et al. Navigating the AI frontier: a primer on the evolution and impact of AI agents[R]. Geneva: World Economic Forum, 2024.
- [6] 中国移动通信有限公司研究院. 人工智能能力与算力网络智慧运营研究与应用[R]. 2024.
- China Mobile Research Institute. Research and application of AI capability and intelligent computing-network operation[R]. 2024.
- [7] 中国移动通信集团有限公司. 算网大脑白皮书 (2022年)[R]. 2022.
- China Mobile Communications Group Co., Ltd. White paper on Computing-Network Brain (2022)[R]. 2022.
- [8] Abou Ali M, Dornaika F, Charafeddine J. Agentic AI: a comprehensive survey of architectures, applications, and future directions[J]. Artificial Intelligence Review, 2025, 59(1): 11.
- [9] Shen Y L, Song K T, Tan X, et al. HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face[C]// Proceedings of the 37th Conference on Neural Information Processing Systems. New York: NeurIPS, 2023: 1–12.
- [10] Guo T C, Chen X Y, Wang Y Q, et al. Large language model-based multi-agents: a survey of progress and challenges[C]//Proceedings of the 33rd International Joint Conference on Artificial Intelligence. New York: ACM Press, 2024: 8048–8057.
- [11] Qian C, Liu W, Liu H Z, et al. ChatDev: communicative agents for software development[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2024: 15174–15186.
- [12] Wang L, Ma C, Feng X Y, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.
- [13] 翟博渊. Ad Exchange 广告交易平台的设计与实现[D]. 北京: 北京交通大学, 2016.
- Zhai B Y. Design and implementation of Ad Exchange platform[D]. Beijing: Beijing Jiaotong University, 2016.
- [14] Huh D, Mohapatra P. Multi-agent reinforcement learning: a comprehensive survey[PP]. arXiv preprint, 2023, arXiv: 2312.10256.
- [15] Wu Q Y, Bansal G, Zhang J Y, et al. AutoGen: enabling next-gen LLM applications via multi-agent conversation[PP]. arXiv preprint, 2023, arXiv:2308.08155.
- [16] Li G H, Hammoud H A A K, Itani H, et al. CAMEL: communicative agents for “mind” exploration of large language model society[PP]. arXiv preprint, 2023, arXiv:2303.17760.
- [17] Roda-Sanchez L, Garrido-Hidalgo C,

Royo F, et al. Cloud - edge microser - vices architecture and service orches - tration: an integral solution for a real - world deployment experience[J]. Inter - net of Things, 2023, 22: 100777.

[18] Shanmugam L, Jangoan S, Sharma K K. Dynamic resource allocation in edge computing for AI/ML applications: ar - chitectural framework and optimization techniques[J]. Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online), 2023, 2(2): 385-397.

[19] Bo X H, Chen X, Dai Q Y, et al. Reflec - tive multi-agent collaboration based on large language models[C]//Proceedings of the Advances in Neural Information. [S.l.:s.n.], 2024: 138595-138631.

[20] Hong S, Zhuge M, Chen J, et al. MetaGPT: meta programming for a multi - agent collaborative framework[PP]. arXiv preprint, 2024, arXiv:2308.00352v7.

[21] Tang X L, Yu H, Li Z X, et al. Multi - session budget optimization for forward auction-based federated learning (MBOS-AFL) [C]//Proceedings of the 42nd International Conference on Ma - chine Learning. PMLR, 2025: 1-14.

[22] Ma C, Li J, Wei K, et al. Trusted AI in multiagent systems: an overview of pri - vacy and security for distributed learn - ing[J]. Proceedings of the IEEE, 2023, 111(9): 1097-1132.

[23] Zhang Z, Yu P, Liang H, et al. A survey on the memory mechanism of LLM - based agents[PP]. arXiv preprint, 2024, arXiv:2404.13501.

[24] Ye R G, Kou W B, Tang M. PraFFL: a preference-aware scheme in fair feder - ated learning[C]//Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1. New York: ACM, 2025: 1797-1808.

[25] Tang Y, Srinivasan U C, James Scott B, et al. End-to-end edge AI service pro - visioning framework in 6G ORAN[C]// Proceedings of the 2025 IEEE 102nd Vehicular Technology Conference. Pis - cataway: IEEE Press, 2026: 1-5.

[26] Shrimal A, Kanagaraj S, Biswas K, et al. MARCO: multi-agent real-time chat orchestration[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Indus - try Track. Kerrville: Association for Computational Linguistics 2024: 1381-1392.

[27] Wang R, Liu J, Chen X, et al. Agentic retrieval-augmented generation: A sur - vey on agentic RAG[PP]. arXiv preprint, 2025, arXiv: 2501.09136.

[28] Sculley D, Holt G, Golovin D, et al. Hidden technical debt in machine learn - ing systems[C]//Proceedings of the 28th Conference on Neural Information Pro - cessing Systems. Red Hook: Curran As - sociates, 2015: 2503-2511.

[29] Wang Z, Moriyama S, Wang W Y, et al. Talk structurally, act hierarchically: a collaborative framework for LLM multi - agent systems[PP]. arXiv preprint, 2025, arXiv:2502.11098.

[30] 王晓哈, 赵鹏, 唐国华, 等. 基于意图增强的算网运营管理系统架构与模型设计[J]. 大数 据, 2025, 11(4): 178-192.

Wang X H, Zhao P, Tang G H, et al. Architecture and model design of opera - tional management system for comput - ing force network based on intent en - hancement[J]. Big Data Research, 2025, 11(4): 178-192.

作者简介



赵鹏（1987-），男，中移九天人工智能科技（北京）有限公司主任研究员、高级工程师，主要研究方向为人工智能、算力网络、5G等。



蔡海涛（1976-），男，中移九天人工智能科技（北京）有限公司解决方案工程师，主要研究方向为人工智能、算网大脑等。



李金翼（1994-），女，中移九天人工智能科技（北京）有限公司解决方案工程师，主要研究方向为人工智能、算网大脑等。



唐国华（1990-），男，中移九天人工智能科技（北京）有限公司解决方案工程师，主要研究方向为人工智能、算力网络、算网大脑等。



史湘荧（2000-），女，北京邮电大学电子工程学院硕士生，主要研究方向为人工智能、多模态大模型等。



袁向阳（1978-），男，中移九天人工智能科技（北京）有限公司部门负责人，主要研究方向为 BSS、OSS 等 IT 支撑系统及 AI 技术在网络智能化中的应用。



王斌（1977-），男，中移九天人工智能科技（北京）有限公司部门负责人，主要研究方向为人工智能大模型、大模型平台和算力基础设施。



邓超（1978-），男，中移九天人工智能科技（北京）有限公司部门负责人，主要研究方向为人工智能、通信网络智能化、大数据和IT技术研发。



冯俊兰（1974-），女，中国移动集团首席科学家，主要研究方向为语音识别、语言理解和数据挖掘等。

收稿日期: 2026-01-04