智能体人工智能互联网：激励相容的分布式组队与工作流

Ya-Ting Yang 和 Quanyan Zhu，纽约大学坦登工程学院电气与计算机工程系，美国纽约州布鲁克林； {yy4348, qz494}@nyu.edu

arXiv:2602.03145v1 [cs.GT] 3 Feb 2026

摘要——大语言模型（LLM）催生了一类新型智能体人工智能系统，它们能够通过调用外部工具进行推理、规划和行动。然而，现有的大多数智能体架构仍是集中式、单体式的，这限制了其可扩展性、专业化和互操作性。本文提出一个可扩展智能体智能框架，称为智能体人工智能互联网。在该框架中，分布于云边基础设施上的自主异构智能体动态组建联盟，以执行任务驱动工作流。我们将智能体协作形式化为一种网络原生模型，并提出一个激励相容的工作流—联盟可行性框架，整合能力覆盖、网络局部性和经济可实施性。为实现可扩展协调，我们建立了最小努力联盟选择问题，并提出一种去中心化联盟形成算法。所提出的框架可作为模型上下文协议（MCP）之上的协调层运行。医疗保健案例研究表明，领域专门化、云边异构性和动态联盟形成能够实现可扩展、稳健且经济可行的智能体工作流。本研究为智能体人工智能互联网这一新兴时代的规范化协调与可扩展性奠定了基础。

索引词—智能体人工智能、工作流、激励相容性、联盟形成

I. 引言

大语言模型（LLM）的出现，以及它们与外部工具、应用程序接口和记忆系统日益紧密的集成，催生了一种新的智能体人工智能范式：这类系统超越被动的信息处理，能够主动推理、规划并行动，以追求高层目标 [1]。当大语言模型嵌入模块化架构时，该范式尤为有效，其中不同组件分别专精于任务规划、工具调用、信息检索、验证和评估。近期研究表明，此类多智能体编排能够实现超越单一单体模型的问题解决能力 [2]–[5]。

这些系统的一项关键赋能技术是模型上下文协议（MCP），它为智能体动态发现和调用外部工具提供了标准化接口。通过抽象工具模式、声明的能力和调用端点，MCP 使工作流能够在运行时进行调整，而不必依赖静态集成。原则上，这种设计支持可扩展性和复用，使智能体能够利用独立于核心模型开发的工具。此外，MCP 也适用于异构的所有权结构：工具可以是专有的、经授权许可的、受访问控制的，也可以是在网络其他位置部署的自主智能体。

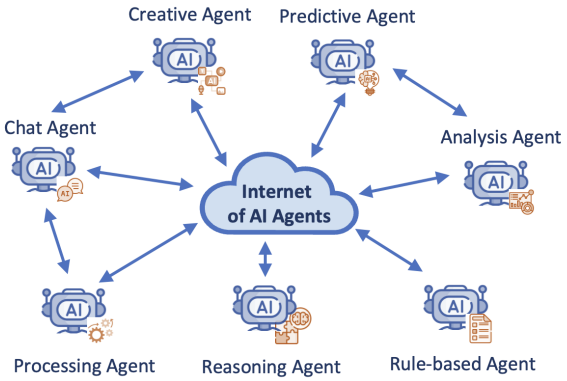


图1：智能体人工智能互联网示意图。每个节点承载一个或多个面向不同任务设计的专用人工智能智能体，并参与共享的通信网络。

尽管接口层面具有灵活性，大多数智能体人工智能系统仍部署在集中且边界明确的计算环境中。虽然这种设计简化了集成并实现了低延迟协调，但从根本上限制了可扩展性。具体而言，可扩展性受到以下因素制约：对集中式所有权与控制的假设、开发和管理多样化专业智能体的组织能力有限，以及不同用户、政策和领域之间互操作性不足。因此，当前系统的扩展主要体现在模型规模或计算能力上，而非功能多样性、组织覆盖范围或能力组合。要开启智能体人工智能发展的下一阶段，就需要从集中编排式部署转向联网的智能体生态系统，让智能体和工具能够跨越组织边界被发现、协调和组合。这一转变将可扩展性重新界定为涉及协调、信任、访问控制和激励的系统级挑战，而不仅仅是计算问题。

为克服集中式智能体人工智能的可扩展性局限，我们提出一种架构愿景：智能体人工智能互联网。在这一框架中，智能体智能并不局限于单一部署或管理域，而是通过分布于云平台、边缘基础设施和机构环境中的分布式节点构成的网络涌现。每个节点托管一个或多个专用人工智能智能体，并参与共享的通信网络，如图1所示。节点可以向网络发布任务，例如调查网络安全事件或优化供应链。其他节点会发现该任务，评估其与自身能力的匹配程度，

并动态组建联盟，共同执行任务工作流。任务完成后，将按照预先定义的机制分配奖励。

我们的贡献可概括如下。首先，我们提出了一种智能体协作的网络原生模型，其中任务动态产生，并由能力要求加以描述。智能体通过网络进行发现，并组建临时联盟，跨越组织边界和云边界来执行任务特定的工作流。其次，我们提出了一个激励相容的工作流—联盟可行性框架，将联盟形成与分布式工作流执行相结合，从而刻画在网络约束和预算条件下联盟何时可行、工作流何时可执行。第三，我们研究最小努力联盟选择问题，并提出一种仅使用局部网络信息的去中心化联盟形成算法，使其适用于开放生态系统。最后，我们指出，所提出的框架可作为协调层，在执行前确定联盟构成、工作流组装和奖励可行性，从而使本研究成为现有 MCP基础设施的一种可扩展、具备激励意识的扩展。

II. 相关工作

多智能体系统中的联盟形成 [6] 深深植根于合作博弈论 [7]。Shapley [8] 提出的可转移效用（TU）博弈经典模型，以及 Chalkiadakis等人 [9]等人在综述中对其所作的进一步拓展，为理解智能体如何协调以实现集体效用并分配奖励提供了基础工具。Shapley值、核心和联盟稳定性等概念已广泛应用于分布式系统 [10]。然而，识别理想的联盟结构在计算上颇具挑战，因为可能的联盟空间会随智能体数量增加而扩大。为此，研究者提出了广泛的算法方法，以计算可行且效用高的联盟 [11]。近期研究也探讨了在通信成本和资源约束等实际限制下形成联盟所面临的挑战 [11], [12]。图约束联盟博弈 [13] 专门将拓扑限制纳入联盟可行性考量，这与我们的研究场景高度相关。我们的研究与这类工作的不同之处在于，我们侧重于组建任务特定联盟，而非优化所有智能体的全局划分。在这些思路的基础上，我们将联盟估值与依赖工作流的结果相结合，从而更紧密地整合网络结构、激励和任务可行性。

去中心化人工智能生态系统中的激励相容性已成为一个活跃且迅速发展的研究领域，尤其是在联邦学习和分布式学习的背景下 [14], [15]。随着学习和决策过程日益依赖多个自主参与者的贡献，如何使个体激励与系统级目标保持一致，已成为一项核心挑战。因此，大量研究致力于设计能够

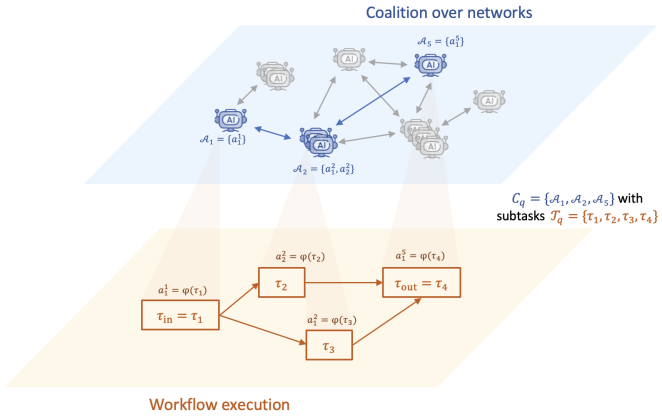


图2：面向特定任务的能力覆盖联盟形成与分布式工作流执行示意图。

鼓励真实参与、公平奖励分配以及自利型智能体持续合作的机制。例如，Ghosh和Hummel [16]研究了智能体任务参与的真实机制；Zhan等人 [17] 分析了兼顾通信成本和智能体贡献的奖励分配方案；Qi等人 [18]则采用基于区块链的框架，通过智能合约实现透明的智能体注册、可验证的任务分配和信誉追踪。

智能体人工智能架构的兴起，得益于大语言模型（LLM）的能力，为分布式推理和自主协调开辟了新范式。早期的模块化智能体系统，如AutoGPT[1] 和BabyAGI[2] ，已展示了将大语言模型串联用于规划和任务执行的强大能力。最近，CAMEL [19]，、AgentVerse [20]，和多智能体辩论 [21]等框架利用大语言模型，将协作性与竞争性的多智能体行为形式化。然而，这些工作大多假定采用集中式编排。相比之下，我们的架构引入了一种去中心化替代方案，其中智能体能力分布在网络中，并通过激励相容的联盟组建实现协调。这与互联网原生智能体计算 [22] 的愿景相一致，并支持可扩展、可互操作的人工智能工作流。

III. 网络上的联盟

我们将分布式多智能体系统视为一个通信网络，并用无向图$\mathcal{G} = (\mathcal{V}, \mathcal{E})$对其建模，其中 $\mathcal{V} = \{1, 2, \ldots, N\}$ 表示节点集合，$\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ 表示节点之间的通信边集合。每个节点 $i \in \mathcal{V}$ 对应一个自主平台，例如计算单元、机构或信息物理系统，这些平台在本地托管一组人工智能智能体，记作 $\mathcal{A}_i = \{a_1^i, a_2^i, \ldots, a_{n_i}^i\}$。每个智能体 $a \in \mathcal{A}_i$都具备一组功能能力 $C(a) \subseteq \mathcal{T}$，其中 $\mathcal{T} = \{c_1, c_2, \ldots, c_M\}$ 定义了全局能力空间，涵盖视觉

[1]https://github.com/Significant-Gravitas/Auto-GPT [2] https://babyagi.org/

识别、摘要生成、逻辑推理或特定领域推理等能力。

A. 领域专门化与任务要求

系统体现出领域专门化。每个节点 i 都与一个子集 $\mathcal{T}_i \subseteq \mathcal{T}$相关联，该子集用于表征其领域专长。假设托管于节点 i 的智能体主要从 $\mathcal{T}_i$获取能力，这反映了实际部署中人工智能模型会针对机构情境进行定制，例如影像中心的放射学推理智能体，或合规中心的法律文档解释器。这种异质性有助于实现模块化和可扩展性，因为节点只需维护与自身专业化相匹配的模型，并在需要时依靠网络提供其他能力。

任务被建模为能力要求，并在节点上动态产生。任务 q 由节点 $i_0 \in \mathcal{V}$发起，其规格由所需能力的多重集表示：$\mathcal{R}_q = \{c_1^{(r_1}\ c_2^{(r_2}\ \dots\ c_M^{(r_M}\}\ r_k \in \mathbb{N}$ r_k ），），，，表示完成该任务需要具备能力 $c_k \in \mathcal{T}$的智能体数量。若发起节点 i_0 无法利用自身智能体池 $\mathcal{A}_{i_0}$满足 $\mathcal{R}_q$，则会启动联盟形成流程，以招募外部支持。联盟的定义是节点 $\mathcal{C}_q \subseteq \mathcal{V}$的一个子集，并满足约束条件 $i_0 \in \mathcal{C}_q$；来自 $\mathcal{C}_q$ 中各节点的智能体集合足以满足任务 q的能力要求。

这一表述为分布式能力发现和协作式任务执行奠定了基础。在这一过程中，领域专业化的节点动态组成临时联盟，共同完成超出单个节点能力范围的任务。

定义*1*（能力覆盖联盟）：给定一个在节点处发起的任务 q $i_0 \in \mathcal{V}$ $\mathcal{R}_q =$，其 $\{c_1^{(r_1}\ c_2^{(r_2}\ \dots\ c_M^{(r_M}\}\ r_k \in \mathbb{N}$ $\mathcal{C}_q \subseteq \mathcal{V}$ 能力要求为），），，若一个节点子集满足以下条件，则称其为能力覆盖联盟：i) $i_0 \in \mathcal{C}_q$ （即发起节点包含在联盟中）；ii) 对于每项所需能力 $c_k \in \mathcal{T}$， $\mathcal{C}_q$ 中所有节点所拥有该能力 c_k 的智能体数量满足：

$$\Big|\{a \in \bigcup_{j \in \mathcal{C}_q} \mathcal{A}_j \mid c_k \in C(a)\}\Big| \geq r_k.$$

定义 *2*（*k*度可行性）：给定通信网络 $\mathcal{G} = (\mathcal{V}, \mathcal{E})$，若存在一个定义1所述的能力覆盖联盟 $\mathcal{C}_q \subseteq \mathcal{V}$，使得每个节点 $j \in \mathcal{C}_q$ 均满足$d(i_0, j) \leq k$，则称在节点 $i_0 \in \mathcal{V}$ 发起、能力要求多重集为 $\mathcal{R}_q$ 的任务 q 具有k度可行性。其中， $d(i_0, j)$ 表示网络 $\mathcal{G}$中的最短路径距离。如果在 i_0的 k跳范围内不存在这样的联盟，则称任务q 为 k度不可行。

定义 *3*（可行性半径）：给定在节点 $i_0 \in \mathcal{V}$ 启动的任务 q，可行性半径 $\rho_q \in \mathbb{N} \cup \{\infty\}$ 定义为满足以下条件的最小整数 k：存在定义 2 中所定义的 k 度可行联盟 $\mathcal{C}_q \subseteq \mathcal{V}$。如果在任何有限距离内都不存在这样的联盟，则令 $\rho_q = \infty$，并称该任务在网络上全局不可行。

B. 分布式工作流执行

一旦形成联盟 $\mathcal{C}_q \subseteq \mathcal{V}$ 以完成任务 q，参与智能体便会协同执行分布式工作流，如图2所示。任务 q 被分解为一组相互依赖的子任务 $\mathcal{T}_q = \{\tau_1, \tau_2, \dots, \tau_L\}$，其中每个子任务 $\tau_\ell \in \mathcal{T}_q$ 都需要特定能力 $c(\tau_\ell) \in \mathcal{T}$。执行结构由有向无环图（DAG）建模，记作 $\mathcal{W}_q = (\mathcal{T}_q, \mathcal{D}_q)$，其中边缘端集合 $\mathcal{D}_q \subseteq \mathcal{T}_q \times \mathcal{T}_q$ 编码了优先关系：若 $(\tau_i, \tau_j) \in \mathcal{D}_q$，则 τ_i 的输出将作为 τ_j 输入的一部分。每个子任务 $\tau \in \mathcal{T}_q$ 都分配给一个智能体 $a = \varphi(\tau) \in \bigcup_{j \in \mathcal{C}_q} \mathcal{A}_j$，以满足能力要求，即 $c(\tau) \in C(a)$。分配函数定义为 φ：$\mathcal{T}_q \mapsto \bigcup_{j \in \mathcal{C}_q} \mathcal{A}_j$。每个智能体 $a \in \bigcup_{j \in \mathcal{C}_q} \mathcal{A}_j$ 执行由元组 (I_a, u_a, f_a, O_a) 描述的局部计算，其中 $I_a \subseteq \mathcal{X}_a$ 是来自上游智能体的结构化输入，$u_a \in \mathcal{U}_a$ 是局部控制输入（例如努力程度或配置），而 f_a： $\mathcal{X}_a \times \mathcal{U}_a \mapsto \mathcal{Y}_a$ 是局部处理函数。输出 $O_a \in \mathcal{Y}_a$ 计算为 $O_a = f_a(I_a, u_a)$。

定义*4*（终端子任务）：若子任务 $\tau \in \mathcal{T}_q$ 没有出边，则称其为终端子任务，即 $\forall \tau' \in \mathcal{T}_q$, $(\tau, \tau') \notin \mathcal{D}_q$。终端子任务的集合记为

$$\mathcal{T}_q^{\text{term}} := \{\tau \in \mathcal{T}_q \mid \nexists \tau' \in \mathcal{T}_q \text{ such that } (\tau, \tau') \in \mathcal{D}_q\}.$$

C. 组合式工作流动力学

分布式工作流 $\mathcal{W}_q = (\mathcal{T}_q, \mathcal{D}_q)$诱导局部智能体计算的递归组合。每个智能体$a = \varphi(\tau)$应用函数 f_a：$\mathcal{X}_a \times \mathcal{U}_a \to \mathcal{Y}_a$，生成输出 $O_a = f_a(I_a, u_a)$，其中 $I_a = (O_{\varphi(\tau')})_{\tau' \in \text{Pred}(\tau)}$聚合前驱任务的输出。前驱集合定义为 $\text{Pred}(\tau) := \{\tau' \in \mathcal{T}_q \mid (\tau', \tau) \in \mathcal{D}_q\}$。由此得到递归关系

$$O_{\varphi(\tau)} = f_{\varphi(\tau)}\left(\left(f_{\varphi(\tau')}(\cdot, u_{\varphi(\tau')})\right)_{\tau' \in \text{Pred}(\tau)}, u_{\varphi(\tau)}\right),$$

其从源子任务展开。对于满足$\text{Pred}(\tau) = \emptyset$的此类子任务$\tau$，输入为外生信号$x_\tau \in \mathcal{X}_{\varphi(\tau)}$，并记为 $O_{\varphi(\tau)} = f_{\varphi(\tau)}(x_\tau, u_{\varphi(\tau)})$。全局任务输出通过对终端输出进行聚合得到：$O_q = \text{Agg}\left(\{O_{\varphi(\tau)} \mid \tau \in \mathcal{T}_{\text{term}_q}\}\right)$，其中Agg是任务特定的聚合算子（例如拼接、求和）。该结构使分布式任务执行具备可追溯性、模块化以及基于原理的性能分析。

IV. 激励相容性与联盟博弈

我们将智能体人工智能团队在网络化系统中的协作建模为一种与动态出现的任务相关联的基于联盟的博弈。令 $q \in \mathcal{Q}$ 表示在时间 t出现的任务，令 $\mathcal{C}_q \subseteq \mathcal{V}$ 表示由为完成任务 q而组建的、具备所需能力的节点联盟（定义见定义1）。每个节点 $i \in \mathcal{C}_q$ 由一个智能体团队 $\mathcal{A}_i = \{a_{i1}, a_{i2}, \dots, a_{in_i}\}$组成，每个智能体 $a \in \mathcal{A}_i$贡献一定的努力 $u_a^q \in \mathbb{R}_{\geq 0}$。因此，节点 i 的总努力为 $u_i^q = \sum_{a \in \mathcal{A}_i} u_a^q$。此外，所有参与节点的集体（合并）智能体努力概况

在 $\mathcal{C}_q$ 中记作$\mathbf{u}^q = (u_a^q)^{a\in\bigcup_{i\in\mathcal{C}_q}\mathcal{A}_i}$。合并后的智能体努力通过任务结果函数 $O_q = f_q(\mathbf{u}^q)$决定任务特定结果，由此产生的任务奖励为（设计为） $R_q = r_q(O_q) \in \mathbb{R}_{\geq 0}$。每个节点 $i \in \mathcal{C}_q$ 承担 i) 努力成本 $c_i(u_i^q)$，即该节点 i上智能体层级努力的总成本；以及 ii) 通信成本 $C_i^{\text{comm}} = \sum_{j\in\mathcal{C}_q\setminus\{i\}} \gamma_{ij}^q$，其中 γ_{ij}^q 表示节点 i 与节点 j之间的通信成本。令 $w_i^q \in \mathbb{R}_{\geq 0}$ 表示分配给节点 i的奖励，并满足预算平衡约束 $\sum_{i\in\mathcal{C}_q} w_i^q = R_q$。那么，节点 i 的净效用为

$$\pi_i^q = w_i^q - c_i(u_i^q) - C_i^{\text{comm}}. \tag{1}$$

根据节点的效用，如果联盟 $\mathcal{C}_q \subseteq \mathcal{V}$ 中每个节点获得的效用都非负，我们就称该联盟具有个体理性（IR）：

$$\pi_i^q \geq 0 \quad \forall i \in \mathcal{C}_q. \tag{2}$$

如果留在当前联盟 $\mathcal{C}_q$ 中所获得的效用至少与其他备选方案一样高，则称其具有激励相容性（IC）：

$$\pi_i^q \geq \pi_i^{\text{out}} \quad \forall i \in \mathcal{C}_q, \tag{3}$$

其中，π_i^{out} 表示退出当前联盟或转而加入其他联盟所获得的效用。在后续分析中，我们假设任务 q 是时刻 t 唯一的任务，并据此处理 $\pi_i^{\text{out}} = 0,\ \forall i \in \mathcal{C}_q$；因此，不参与 $\mathcal{C}_q$ 所获得的效用为零。在这种情况下，IC 约束等价于 IR 约束。多个任务同时出现（从而形成多个联盟和 $\pi_i^{\text{out}} \neq 0$）的情形留待未来研究。

A. 联盟与工作流之间的一致性

可以看出，联盟形成、工作流执行和奖励设计之间存在根本性的耦合关系。工作流的结果决定任务奖励总额，而任务奖励总额又决定联盟的可行性（个体理性和激励相容性）。反过来，联盟结构（例如参与智能体及其分配的努力）决定工作流能否成功执行。当这种相互依赖关系得到满足时，我们称系统具有工作流—联盟可行性：根据联盟结构，工作流能够执行，且由此产生的奖励能够为所有联盟成员提供可实施的激励机制。

定义 *5*（工作流—联盟可行性）：设 q 为在节点 $i_0 \in \mathcal{V}$ 发起的任务，能力要求为 $\mathcal{R}_q$，奖励函数为 r_q。相关工作流 $\mathcal{W}_q = (\mathcal{T}_q,\ \mathcal{D}_q)$ 由相互依赖的子任务 $\tau \in \mathcal{T}_q$组成，每个子任务都需要由分配至该子任务的智能体 $a = \varphi(\tau)$付出努力 u_a。设 $\mathcal{C}_q \subseteq \mathcal{V}$ 为一个拟议的节点联盟，并设 $\mathcal{A}_q = \bigcup_{i\in\mathcal{C}_q} \mathcal{A}_i$ 表示可用于任务 q的智能体集合。若满足以下条件，则称这一对 $(\mathcal{W}_q,\ \mathcal{C}_q)$ 在工作流 — 联盟意义下可行：

(ii) 能力覆盖： $\mathcal{C}_q$ 是定义1中所定义的能力覆盖联盟。

(ii) 能力一致： 存在一个分配 $\varphi{:}\mathcal{T}_q \mapsto \mathcal{A}_q$，使得对于每个 $\tau \in \mathcal{T}_q$，被分配的智能体$a = \varphi(\tau)$都满足所需的能力：$c(\tau) \in C(a)$。

（iii）**定义良好的工作流输出：** 由努力 $\{u_a\}_{a\in\mathcal{A}_q}$驱动的子任务函数 $f_{\varphi(\tau)}$递归应用后，会产生定义良好的全局结果 $O_q \in \mathbb{R}$。也就是说，$O_{\varphi(\tau)} = f_{\varphi(\tau)}\big(I_{\varphi(\tau)},\ u_{\varphi(\tau)}\big)$，其中终端输出汇总为 $O_q = \text{Agg}\big(\{O_{\varphi(\tau)}\}_{\tau\in\mathcal{T}^{\text{term}}_q}\big)$。

(iv) 奖励可实现性： 由工作流结果 $R_q = r_q(O_q)$ 产生的任务奖励有限且非负：$0 \leq R_q < \infty$。**(v) 预算可行性：** 联盟成员产生的运营和通信总成本可由任务奖励覆盖：$\sum_{i\in\mathcal{C}_q}\big[c_i(u_i) + C_i^{\text{comm}}\big] \leq R_q$。**(vi) 激励相容性：** 存在一个奖励分配 $\{w_i^q\}_{i\in\mathcal{C}_q}$，满足 $w_i^q \geq c_i(u_i) + C_i^{\text{comm}}$，$\forall i \in \mathcal{C}_q$，其中 $\sum_{i\in\mathcal{C}_q} w_i^q = R_q$。

定义 *6*（k度工作流—联盟可行性）：令 $G = (\mathcal{V},\ \mathcal{E})$ 表示通信图，令 $d(i,\ j)$ 表示节点 i 与 j 在 G 中的最短路径距离。若存在定义 5 中所定义的工作流联盟可行对 $(\mathcal{W}_q,\ \mathcal{C}_q)$，使得 $d(i_0,\ j) \leq k,\ \forall j \in \mathcal{C}_q$，则称任务 q 为 k 度工作流—联盟可行。对于给定的 k，若不存在这样的联盟，则称任务 q 为 k 度不可行。若不存在这样的 $k \in \mathbb{N}$，则称任务 q 在智能体人工智能互联网中全局不可行。

B. 最小努力联盟

在本节中，我们旨在确定一个支持给定工作流的联盟，同时尽量减少参与智能体付出的总努力。

定义 *7*（最小努力联盟选择）：设 q为一个任务，其关联的工作流为 $\mathcal{W}_q = (\mathcal{T}_q,\ \mathcal{D}_q)$，并在节点 $i_0 \in \mathcal{V}$处启动。对于给定的跳数半径 k，定义

$$\mathcal{C}_q^{(k)} \triangleq \Big\{\mathcal{C}_q \subseteq \mathcal{V} \mid \mathcal{C}_q \text{ satisfies Definition 6}\Big\}. \tag{4}$$

最小努力联盟选择问题旨在寻找一个联盟 $\mathcal{C}_q^\star \in \mathcal{C}_q^{(k)}$，使执行该工作流所需的智能体总努力最小：

$$\mathcal{C}_q^\star \in \arg\min_{\mathcal{C}_q\in\mathcal{C}_q^{(k)}} \sum_{i\in\mathcal{C}_q}\sum_{a\in\mathcal{A}_i} u_a. \tag{5}$$

如果 $\mathcal{C}_q^{(k)} = \emptyset$，则在跳数半径 k内不存在可行联盟。此时，系统可以逐步增大 k，直至找到可行联盟或达到预定义的探索上限。

工作流联盟形成算法（如算法1所述）提出了一种去中心化搜索与选择过程：智能体在节点i_0处发起任务，并在不断扩大的跳数半径邻域内逐步发现和评估候选联盟。在每个半径 k内，算法枚举满足 k-度可行性局部性约束（定义 2）的联盟 $\mathcal{C}_q \subseteq \mathcal{N}^{(k)}(i_0)$。对于每个候选联盟，算法尝试分配

算法1： 工作流联盟形成

输入： 在节点 i_0启动的任务 q，工作流 $\mathcal{W}_q = (\mathcal{T}_q, \mathcal{D}_q)$，能力要求 $\mathcal{R}_q$，$\{c(\tau_\ell)\}_{\ell=1}^{L}$，最大搜索半径 $k_{\max}$

输出： 可行联盟 $\mathcal{C}^\star_q$ 和工作流分配，或不可行

初始化 $k \leftarrow 1$;

当 $k \le k_{\max}$ **do**

  计算节点 i_0 的 k 跳邻域：$\mathcal{N}^{(k)}(i_0) := \{i \in \mathcal{V} \mid d(i, i_0) \le k\}$。

  为 $\mathcal{R}$ 构造候选联盟集合 $\mathcal{F}_q^{(k)}$：$\mathcal{F}_q^{(k)} := \{\mathcal{C}_q \subseteq \mathcal{V} \mid \mathcal{C}_q$ 满足定义 2$\}$。

  若 $\mathcal{F}_q^{(k)} = \emptyset$ **则**

    $k \leftarrow k+1$; **继续**

  初始化 $\mathcal{C}_q^{(k)} \leftarrow \emptyset$;

  对于每个 $\mathcal{C}_q \in \mathcal{F}_q^{(k)}$ **do**

    若 $(\mathcal{W}_q, \mathcal{C}_q)$ 满足工作流—联盟可行（定义 5）带有分配 φ**则**

      将 $\mathcal{C}_q$ 添加到 $\mathcal{C}_q^{(k)}$;

  如果 $\mathcal{C}_q^{(k)} = \emptyset$ **则**

    $k \leftarrow k+1$; **继续**

  选择 $\mathcal{C}_q^\star$ 如定义 7 中所述：$\mathcal{C}_q^\star \in \arg\min_{\mathcal{C}_q \in \mathcal{C}_q^{(k)}} \sum_{i \in \mathcal{C}_q} \sum_{a \in \mathcal{A}_i} u_a$.

  返回 $(\mathcal{C}_q^\star, \mathcal{W}_q)$;

返回 不可行

工作流子任务，并以符合能力的方式分配给联盟成员，同时验证所得分配是否满足工作流—联盟可行性（定义 5），包括奖励可实现性、预算可行性和激励相容性。

如果至少识别出一个可行联盟，算法会选择使智能体总努力最小的联盟，从而解决最小努力联盟选择问题（定义 7）。如果当前跳数半径内不存在可行联盟，则将搜索半径扩大至 $k+1$，直到找到可行解或达到预先设定的最大探索半径$k_{\max}$。这一过程将工作流规划与去中心化联盟形成紧密结合，确保新出现的任务由在结构上可行且在经济上可实施的联盟提供支持，从而实现智能体人工智能互联网。

V. 案例研究：医疗保健领域的智能体人工智能互联网

智能体人工智能互联网在协作医疗保健领域有着极具吸引力的应用场景 [23]。在区域医疗生态系统中，基层医疗诊所、诊断影像中心、专科医院、远程医疗服务提供商和保险平台等机构，会部署与各自机构职能相匹配的异构人工智能能力。这些能力分布在边缘端和云基础设施的组合环境中：诊所和远程医疗服务提供商通常会在边缘端托管轻量级、对延迟敏感的智能体，

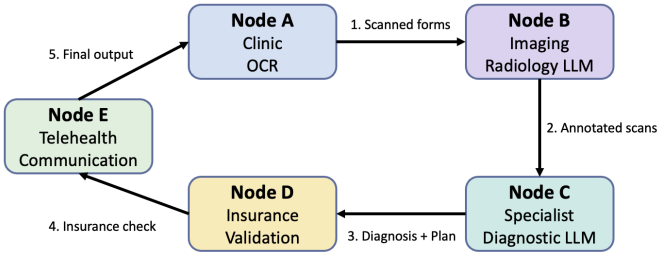


图 3：具有方向性信息流的智能体医疗工作流。系统形成一个闭环，其中下游诊断、验证和通信输出会返回诊所，以便进行协调和最终确认。

而影像中心、医院和保险公司则可能在云端运行计算密集型的基于大语言模型的智能体，这些智能体通常使用放射学报告、药学记录和临床指南等领域专属数据进行微调。

没有任何单一机构拥有处理复杂多阶段患者病例所需的全部能力。专业知识、数据可用性、隐私约束和计算资源方面的差异，使集中式解决方案难以切实可行。智能体人工智能互联网使这些机构能够组成动态的任务驱动联盟，整合网络中互补的能力。

A. 医疗工作流

图 3 展示了一个具有代表性的分布式智能体医疗工作流，仅用于说明。患者前往一家小型基层医疗诊所（记为节点 A），提供手写接诊表、放射影像和既往化验结果，同时还存在需要专门评估的模糊症状。节点 A 配备了能够进行光学字符识别（OCR）和结构化表单解析的智能体，但往往缺乏先进的诊断能力。

认识到这些局限后，节点 A 通过与其他节点组成联盟，启动分布式工作流。如图 3 的步骤 1 所示，节点 A 的OCR智能体将接诊表数字化并提取结构化信息，随后将信息转发至节点B——一家设有基于云端的放射学智能体的区域影像中心。这些由大语言模型驱动的智能体会分析医学影像、生成标注扫描图像，并精确定位异常（步骤 2）。

工作流进入节点C，即一家专科医院，其云端部署的诊断大语言模型执行鉴别诊断、对候选病症进行排序，并在需要时建议额外检查（步骤3）。借助领域特定的临床规程，这些智能体提供情境感知评估，其可靠性超过通用模型。

综合得出初步诊断和治疗方案后，节点 A 会咨询节点D——保险与药房集成平台。节点D的智能体会核实保险覆盖范围、筛查药物相互作用，并确认药物供应情况（步骤4）。最后，节点E（远程医疗服务提供方）会生成面向患者的说明和知情同意文件，并安排随访咨询，以此完成工作流，

然后将最终输出返回给节点 A，以便进行协调和交付（步骤 5）。

B. 数值设置

根据图 3 中的工作流，五个子任务映射到五个机构：

τ_1：诊所 (A) 的 OCR 和接诊，τ_2：影像中心 (B) 的放射学解读 (RAD)，τ_3：专科医院 (C) 的鉴别诊断 (DX)，τ_4：保险提供商 (D) 的保单验证 (VAL)，τ_5：远程医疗 (E) 的远程会诊 (CONS)。

由于智能体与节点一一映射，令每个 ℓ对应 $u_i = u_\ell$ 和 $i = i_\ell$。每个节点都会产生运营成本 $c_i(u_i)$ 和通信开销 $C_i^{\text{comm}} \geq 0$。平台必须分配奖励份额 $\{w_i^q\}_{i\in\mathcal{V}}$，以同时满足激励约束和预算约束。

1) 智能体努力（归一化单位）：对于每个节点 $i \in \mathcal{V}$，令 $u_i \in \mathbb{R}_{\geq 0}$ 表示在节点i付出的智能体努力。变量 u_i 以归一化智能体努力单位计量，其中一个单位对应一组固定的智能体计算与交互活动，例如一个推理—执行周期、一个工具调用轮次，或固定数量的推理令牌。具体而言，u_i 汇总了节点 i所托管智能体的审议深度、交互频率、规划复杂度和推理负载。这种归一化方式使异质的智能体活动能够在统一尺度上表示，同时保留其内在的成本—质量权衡。形式上，$[u_i]$ = 归一化智能体努力单位；如有需要，可通过固定的缩放系数还原令牌、API 调用或 GPU 秒数等实际数量。

2) 智能体审议有效性：令 $g_i : \mathbb{R}_{\geq 0} \mapsto [0, 1]$ 表示节点 i 处的智能体审议有效性，将努力映射为任务阶段可靠性的归一化度量。我们假设 g_i 单调递增、连续可微且为凹函数，以体现额外审议带来的边际收益递减。在数值研究中，我们采用

$$g_i(u_i) = 1 - e^{-\rho_i u_i}, \ \rho_i > 0, \tag{6}$$

其中，ρ_i 量化审议效率，即单位智能体努力带来的可靠性边际提升。

3) 工作流输出：给定链式医疗工作流，端到端任务输出定义为

$$O_q(u) = \prod_{i\in\mathcal{V}} \alpha_i \, g_i(u_i), \tag{7}$$

其中 $\alpha_i \in (0, 1]$ 表示节点 i处人工智能系统或机构流程的基线可靠性。乘法结构反映了误差在医疗流程链各连续阶段中的累积，例如OCR、放射学、诊断、验证和会诊。

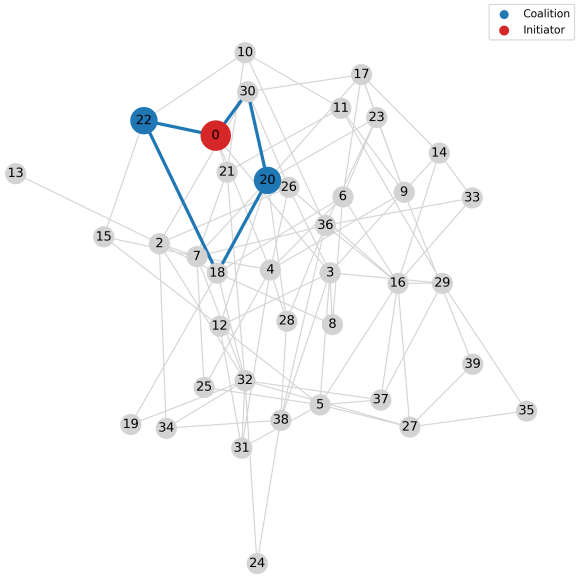


图 4：可行工作流联盟示例。发起节点以红色显示，选定的联盟成员以蓝色显示。高亮边缘端表示用于协调工作流执行的联盟诱导通信路径。

4) 奖励函数：任务奖励为giv 由以下公式给出

$$R_q(u) = \beta \log(1 + O_q(u)), \qquad \beta > 0, \tag{8}$$

其中，R_q 以归一化效用单位衡量，代表风险调整后的医疗价值（例如报销、避免的伤害或责任减免）。对数形式体现了决策置信度提高所带来的边际价值递减，并反映了实际报销上限。

5）智能体执行成本：每个节点 $i \in \mathcal{V}$ 都会产生一项成本 c_i：$\mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$，其建模方式如下

$$c_i(u_i) = \kappa_i^{\text{cpu}} \, u_i + \kappa_i^{\text{lat}} \, u_i^2, \tag{9}$$

其中，κ_i^{cpu}（单位努力的效用）表示直接计算成本，例如 GPU 使用或 API 定价；κ_i^{lat}（单位努力平方的效用）则体现延迟放大、同步以及随审议深度超线性增长的编排开销。每个节点还会产生固定的通信成本 $C_i^{\text{comm}} \geq 0$，涵盖智能体间消息传递、安全数据交换以及机构协调开销。

C. 工作流—联盟形成

我们考虑一个包含 $N = 40$ 个节点的无向埃尔德什–雷尼图，其中每个节点代表一项智能体人工智能服务，并配备从 {OCR、RAD、诊断、VAL、CONS} 中随机抽取的医疗相关能力子集。每个节点还根据不同的智能体特征进行参数化，包括审议效率、线性计算成本、二次延迟惩罚、固定协调开销和基准努力水平，具体设定见第 V-B 节的智能体成本与质量模型。

需要全部五项能力的工作流在节点$i_0 = 0$处发起。联盟搜索与选择算法（算法

1）逐步扩大跳数半径，探索邻域，直至最大值 $k_{\max} = 4$，并枚举包含发起节点、满足能力覆盖且符合工作流联盟经济可行性条件（例如预算/奖励可行性以及简化的激励相容性/个体理性）的候选联盟。

在图 4 所示的运行中，该算法在跳数半径 $k = 2$ 处识别出一个可行联盟。所选联盟由节点 $\{0, 20, 22\}$ 组成。根据实际实现的能力分配，节点 0 提供 {CONS,OCR}，节点 20 提供 {诊断,RAD}，节点 22 提供 {VAL}，共同覆盖所有所需的工作流能力。在搜索过程中遇到的所有可行联盟中，这个三节点联盟实现了最低的总智能体执行成本。该图还表明，尽管网络中存在许多能力重叠的备选节点，一个规模较小、位置合适且局限于适度跳数半径内的联盟仍可实现完整的工作流可行性。

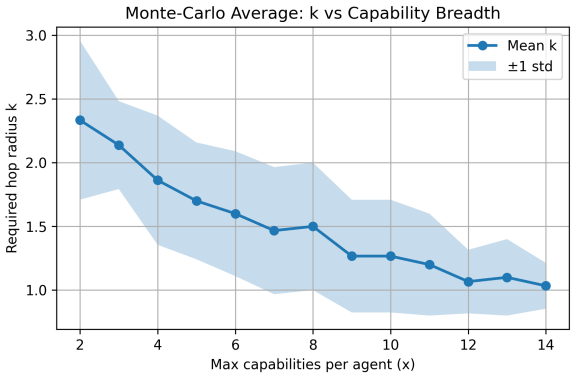
所选联盟的经济效率由总成本 4.51 及相应的工作流奖励 8.25 加以量化，从而产生严格为正的盈余。最佳可行联盟成本在搜索过程中的演变如图5c所示。该曲线先急剧下降，随后趋于稳定，表明近似最优可行联盟在搜索早期即被发现，后续评估主要是确认最优性，而非带来进一步改进。

除了这一单次实现之外，我们还通过蒙特卡洛平均考察了智能体能力广度对协调需求的影响。图5a报告了实现工作流可行性所需的平均跳数半径随每个智能体的最大能力数量变化的情况。随着能力广度增加，所需的协调半径降低，揭示了局部智能体丰富度与网络级协调之间的替代效应。同样，图5b显示，平均联盟规模随着能力广度增加而减小，证实了这样一种直觉：能力更丰富的智能体降低了联盟扩张的需求。

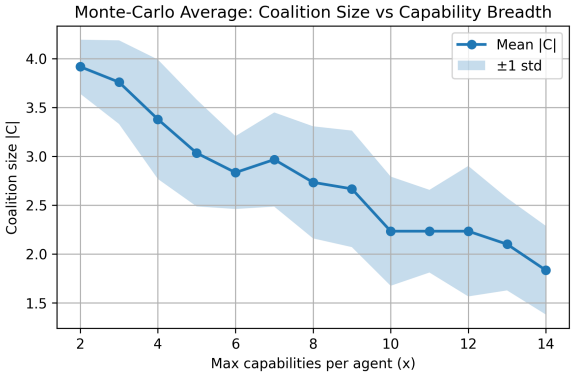
六、讨论：作为 MCP 的扩展

模型上下文协议（MCP）通过公开外部工具的模式、声明的能力和调用端点，为发现和调用外部工具提供了标准化且注重安全的接口。本文开发的工作流联盟算法与 MCP 兼容，并不修改其执行语义。相反，它严格运行在 MCP 之上，作为执行前协调与规划层。我们将这一增强架构称为 C+MCP，其中 MCP 仍作为执行基底，而 C 表示负责联盟形成、可行性评估和成本感知工具选择的协调层。

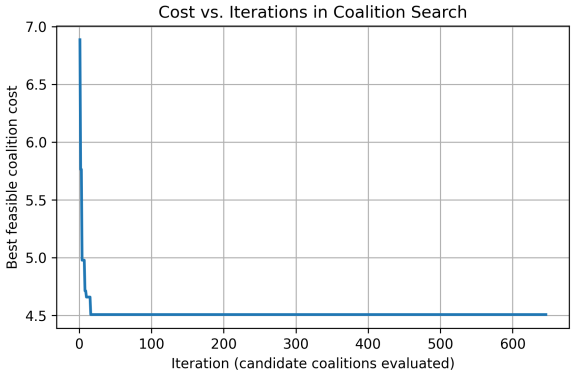
在 C+MCP 的诠释中，每个 MCP工具都被建模为一个智能体实体，它不仅会公布自身的功能能力，还会公布依赖于投入的质量、计算成本、延迟和协调开销等抽象描述。发起方大语言模型充当领导节点 i_0，而 MCP工具则构成周围网络中的候选联盟



(a) 平均所需跳数半径 k 与单个智能体的能力广度 x的关系。阴影区域表示 ± 1 标准差。



(b) 平均联盟规模 $|C|$ 与能力广度 x的关系，展示了随着智能体丰富度增加，协调需求随之降低。



(c) 在一次具有代表性的联盟形成运行中，最佳可行联盟成本与搜索迭代次数的关系。

图 5：工作流联盟形成的扩展行为。增加单个智能体的能力广度，会同时降低满足可行性所需的协调半径和联盟规模。就可实现的最佳成本而言，联盟搜索表现出快速收敛。

成员。这些抽象描述丰富了工具选择的决策逻辑，同时无需对 MCP 的接口、模式或运行时行为做任何更改。

具体而言，C+MCP 中的协调发生在工具调用之前。给定工作流规范，领导智能体首先解决工作流—联盟可行性问题：它

在可用工具的有界邻域内进行搜索，以确定一个子集，使其联合能力能够覆盖该工作流，且其潜在聚合奖励超过执行成本与协调成本之和。该过程在算法1中进行了形式化，并得到一个在功能上充分且在经济上可实施的联盟。

在实际运行中，MCP 保持不变。一旦确定可行联盟，领导智能体便会根据推导出的工作流分配，使用标准 MCP 调用来调用每个入选工具。所有模式验证、安全检查和上下文处理均与现有 MCP 部署中的做法完全一致。C$^+$MCP 的创新完全体现在执行前协调层，该层会在调用任何工具之前确定联盟构成、工具参与情况和努力分配。

VII. 结论

本文将基于大语言模型的智能体之间的协调重新定义为网络化系统问题，推动了可扩展智能体智能的发展。我们不再依赖集中式编排，而是提出智能体人工智能互联网这一框架；在该框架中，分布于云边基础设施上的自主异构智能体能够动态组建联盟，以执行特定任务的工作流。通过提出激励相容的工作流—联盟可行性框架和去中心化协调算法，我们展示了可扩展的智能体人工智能如何从专业智能体在部分信息和网络约束下的有原则交互中涌现。

本研究的一个核心观点是，仅通过增加模型规模难以实现智能体人工智能的可扩展性。相反，这需要借助架构和经济机制，使智能体能够相互发现、有效协调，并以激励相容的方式分享奖励。我们的研究结果表明，联盟形成、工作流执行和定价（奖励）本质上是相互耦合的问题，需要共同解决，才能确保在开放的分布式环境中具备运行可行性和经济可行性。所提出的框架为应对这些挑战奠定了基础，同时兼容现有的智能体基础设施。

智能体人工智能互联网开辟了广泛的未来研究方向。亟待解决的挑战包括设计定价和激励机制 [24],基于市场的协调协议、责任归属 [25],以及用于规范独立运营的智能体之间交互的服务级别协议。将信任、声誉和合规纳入联盟形成也仍是一个开放问题，尤其是在医疗保健和网络安全等领域 [26]。

参考文献

[1] D. B. Acharya、K. Kuppan和B. Divya，“智能体人工智能：面向复杂目标的自主智能——一项全面综述”，*IEEEAccess*，2025。

[2] Q. Wu、G. Bansal、J. Zhang、Y. Wu、B. Li、E. Zhu、L. Jiang、X. Zhang、S. Zhang、J. Liu等，“Autogen：通过多智能体对话实现下一代大语言模型应用”，载于第一届语言建模会议，2024。[3] R. Shu、N. Das、M. Yuan、M. Sunkara和Y. Zhang，“迈向有效的生成式人工智能多智能体协作：面向企业应用的设计与评估”，arXiv预印本arXiv:2412.05449，2024。[4] K.-T. Tran、D. Dao、M.-D. Nguyen、Q.-V. Pham、B. O’Sullivan和H. D. Nguyen，“多智能体协作机制：大语言模型综述”，arXiv预印本arXiv:2501.06322，2025。[5] A. X. Tian、R. Zhang、J. Tang、Y. M. Cho、X. Li、Q. Yi、J. Wang、Z. Zhang、D. Qi、Z. Li等，“超越最强大语言模型：基准测试中的多轮多智能体编排与单一大语言模型之比较”，arXiv预印本arXiv:2509.23537，2025。[6] S. Sarkar、M. Curado Malta和A. Dutta，“联盟形成在多智能体系统中的应用综述”，《并发与计算：实践与经验》，第34卷，第11期，e6876页，2022。[7] R. Branzei、D. Dimitrov和S. Tijs，《合作博弈论中的模型》。Springer，2008。[8] L. S. Shapley，“n人博弈的一种值”，第307–317页，1953。[9] G. Chalkiadakis、E. Elkind和M. Wooldridge，《合作博弈论的计算问题》。Morgan and Claypool出版社，2011。[10] T. Sandholm、K. Larson、M. Andersson、O. Shehory和F. Tohm´e，“具有最坏情况保证的联盟结构生成”，《人工智能》，第111卷，第1–2期，第209–238页，1999。[11] T. Rahwan、S. D. Ramchurn和N. R. Jennings，“最优联盟结构生成的任意时间算法”，《人工智能研究杂志》，第34卷，第521–567页，2009。[12] W. Saad、Z. Han、M. Debbah、A. Hjorungnes和T. Bas¸ar，“面向通信网络的联盟博弈论”，《IEEE信号处理杂志》，第26卷，第5期，第77–97页，2009。[13] T. P. Michalak、T. Rahwan、P. Szczepanski和N. R. Jennings，“博弈论网络中心性中Shapley值的高效计算”，《人工智能研究杂志》，第46卷，第607–650页，2013。[14] X. Tu、K. Zhu、N. C. Luong、D. Niyato、Y. Zhang和J. Li，“联邦学习的激励机制：从经济学与博弈论视角”，《IEEE认知通信与网络汇刊》，第8卷，第3期，第1566–1593页，2022。[15] Y. Zhan、J. Zhang、Z. Hong、L. Wu、P. Li和S. Guo，“联邦学习激励机制设计综述”，《IEEE计算新兴主题汇刊》，第10卷，第2期，第1035–1044页，2021。[16] A. Ghosh和P. Hummel，“通过机制设计激励对等网络中的参与”，《管理科学》，第62卷，第7期，第2085–2102页，2016。[17] Z. Zhan、F. Bu、B. Yang等，“学习合作博弈中的公平分配规则”，arXiv预印本arXiv:2002.06142，2020。[18] M. Qi、T. Zhu、L. Zhang、N. Li和W. Zhou，“迈向去中心化大语言模型多智能体系统中透明且激励相容的协作：一种区块链驱动的方法”，arXiv预印本arXiv:2509 16736，2025。[19] Z. Liu、Y. Shen、X. Li等，“CAMEL：用于大规模语言模型社会心智探索的交流智能体”，arXiv预印本arXiv:2305.15002，2023。[20] Y. Ma、Y. Wang、T. Lin等，“AgentVerse：促进多智能体协作并探索涌现行为”，arXiv预印本arXiv:2402.09629，2024。[21] Y. Du、J. Gao、S. Ghosh等，“基于大语言模型的多智能体辩论”，arXiv预印本arXiv:2305.14325，2023。[22] OpenAI，“智能体人工智能：使大语言模型能够自主行动”，2024，https://openai.com/research/agentic-ai。[23] M. Milne-Ives、C. De Cock、E. Lim、M. H. Shehadeh、N. De Pennington、G. Mole、E. Normando、E. Meinert等，“人工智能对话智能体在医疗保健中的有效性：系统综述”，《医学互联网研究杂志》，第22卷，第10期，e20346页，2020。[24] Y.-T. Yang和Q. Zhu，“Pact：一种用于由大语言模型驱动的智能体人工智能服务定价的契约理论框架”，arXiv预印本arXiv:2505.21286，2025。[25] Y. Ge、Y.-T. Yang和Q. Zhu，“人工智能引发事件中的责任归属：一种用于问责的计算反思均衡框架”，arXiv预印本arXiv:2404.16957，2024。[26] T. Li、Y.-T. Yang、Y. Pan和Q. Zhu，“从文本到盾牌：大语言模型与网络安全的融合”，arXiv预印本arXiv:2505.00841，2025。