迈向集体人工智能科学：基于大语言模型的多智能体系统需要从盲目试错转向严谨科学

Jingru Fan[1, *], **Dewen Liu**[2, *], **Yufan Dang**[3], **Huatao Li**[1], **Yuheng Wang**[1], **Wei Liu**[4], **Feiyu Duan**[2], **Xuanwen Ding**[2], **Shu Yao**[1], **Lin Wu**[1], **Ruijie Shi**[3], **Wai-Shing Leung**[3], **Yuan Cheng**[1], **Zhongyu Wei**[2], **Cheng Yang**[5], **Chen Qian**[1, †], **Zhiyuan Liu**[3, †], **Maosong Sun**[3]

[1]上海交通大学, [2]复旦大学 [3]清华大学, [4]伦敦国王学院, [5]鹏城实验室

{qianc, fanjingru510}@sjtu.edu.cn, dwliu23@m.fudan.edu.cn, liuzy@tsinghua.edu.cn

arXiv:2602.05289v1 [cs.CL] 5 Feb 2026

摘要

近年来，大语言模型（大语言模型）的进展极大拓展了多智能体系统（多智能体系统）的能力，并展现出在广泛复杂且开放领域中的显著成效。然而，尽管取得了快速进展，该领域仍严重依赖经验性试错，缺乏系统优化和改进所必需的统一且有原则的科学框架。这一瓶颈源于归因的不明确性：首先，缺乏结构化的因素分类体系，使研究人员只能进行无指导的调整；其次，缺乏统一指标，无法区分真正的协作增益与单纯的资源累积。本文倡导通过综合框架转向设计科学。我们主张建立协作增益指标（Γ）作为科学标准，以将内在增益与预算增加所带来的增益区分开来。借助 Γ，我们提出因子归因范式，以系统识别协作驱动因素。为此，我们构建了系统化的多智能体系统因素库，将设计空间组织为控制层预设与信息层动态。最终，该框架推动了从盲目实验向严谨科学的转变，为迈向真正的集体人工智能科学铺平道路。

1. 引言

随着大语言模型（LLMs）推理能力的增强，具备目标导向规划和任务执行能力的强大自主智能体应运而生（Minaee 等人，2024；Zhou 等人，2025a；Wang 等人，2024b）。通过整合内部记忆和外部工具使用，这些智能体能够独立执行复杂任务，例如代码生成和数学问题求解（Xi 等，2025；Schick 等，2023）。尽管这些智能体本质上受限于各自的能力边界，难以在开放领域中保持全面的专业能力，但受生物群体和人类文明中集体智能成功经验的启发（详见附录 A），多智能体系统（MAS）已成为突破这些限制的一个有前景的方向（Chen 等，2023；Li 等，2024；Guo 等，2024）。通过组织不同个体协同工作，多智能体系统能够超越单个智能体的部分局限，使智能体凭借集体智能解决多轮、多功能且开放式的问题（Tran 等，2025；Maldonado 等，2024）。因此，多智能体系统范式已在科学研究（Ghareeb 等，2025）、软件工程（Qian 等，2024）、基础设施管理（Seff 等，2023）、医疗服务（Liu 等，2025a）、金融分析（Xiao 等，2025）和社会科学（Zhang 等，2024）等复杂且开放的领域取得了显著成功。

然而，尽管多智能体系统领域涌现了大量成功案例，其发展仍依赖工程驱动的盲目试错，缺乏指导系统性改进的科学依据（Pan 等人，2025；Guo 等人，2024）。因此，研究人员通常只能通过经验性调整关键因素来优化性能，例如组织结构（Li 等人，2024）。这种对经验性调整的依赖，本质上源于无法进行严谨的因素归因。尽管此类经验性优化有时能提升性能，但这些提升的真正归因仍不明确。

从多智能体系统构建的角度来看，归因歧义体现为两项关键挑战。首先，在启动多智能体系统优化之前，研究人员必须选择具体因素来验证性能增益；然而，由于缺乏针对庞大因素空间的结构化分类体系，这种选择受到阻碍（Maldonado 等人，2024）。如果没有标准化的候选因素集作为流程的依据，研究人员就只能进行缺乏指导的因素调整，无法系统地考察潜在的协作—

*共同贡献。
†通讯作者。

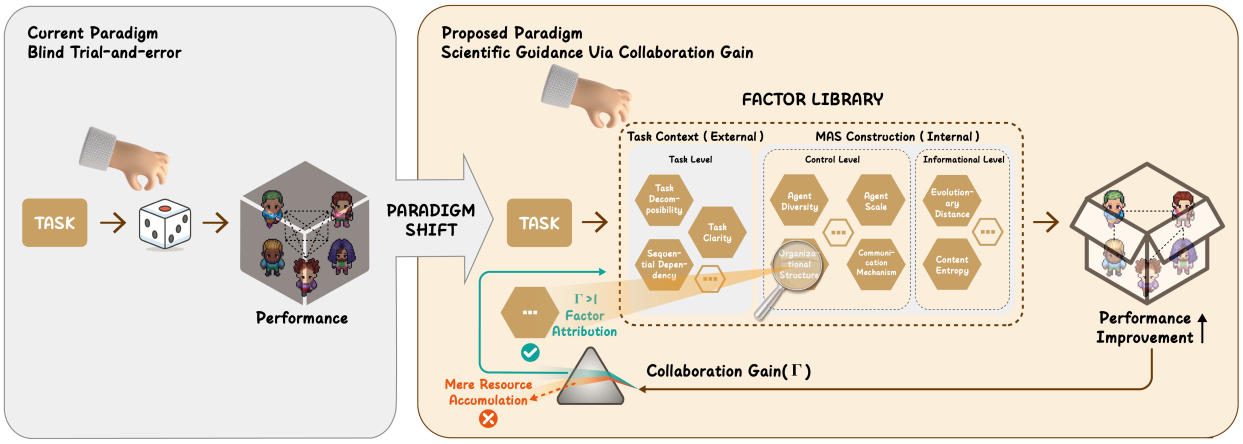


图 *1.* 范式转变：从盲目试错走向科学指导。左侧（当前）：一个不透明的黑盒，性能增益具有随机性且无法归因。右侧（所提出）：一种白盒范式。研究人员从库中选择因素来构建MAS；随后，将观测到的性能输入 Γ （棱镜），以过滤掉单纯的资源累积，从而分离出真正的协作增益。这一分析步骤天然地实现了因素归因。

驱动因素（Kim 等人，2025）。更为关键的是，在价值验证方面，缺乏统一指标，使得人们无法区分某项具体因素的变更究竟带来了真正的协作增益，还是仅仅反映了资源累积（能力增长但没有协作）（Zhang 等人，2025c）。当前针对特定任务的指标本质上混淆了这两种性质迥异的效应：协作带来的内在增益（即智能体协作促成的能力增长），以及仅由计算预算增加带来的性能提升（Qian 等人，2025）。由于既缺乏因素结构化框架，也缺少精确的指标诊断信号，研究人员无法系统地识别驱动真正协作的设计因素（Renney 等人，2026）。这一局限导致大量计算资源浪费在无效设计上；更重要的是，它阻碍了科学原理的建立，使多智能体系统无法发展成为一门严谨的科学（Kapoor 等人，2025）。

为应对这些挑战并指导MAS的科学构建，我们提出了三项核心建议，形成一个整合范式，如图 1所示。认识到科学有效的评估标准是区分有效设计的前提，我们优先提出度量指标和归因方法，该方法随后指导对因素空间的系统探索（Zhu 等人，2026）：

1. **构建协作增益指标（Γ）：** 首先，我们倡议建立协作增益指标（Γ），将其作为系统优化的量化反馈信号。我们将 Γ 定义为在计算资源约束相当的条件下，多智能体系统相对于单智能体系统的性能比率。通过将真正的协作增益与单纯的资源累积分离开来（Du 等人，2025），该指标提供了识别有效优化方向所必需的量化反馈。因此，它也构成了一个有原则依据的诊断信号，为整个科学框架奠定基础。

2. **因子归因范式：** 借助这一度量指标，我们提出设计一种因子归因范式，以识别真正的协作驱动因素。其核心机制遵循两步式的顺序逻辑：我们首先检验修改某一因素是否会带来可测量的性能改进。只有在观察到此类改进后，我们才应用 Γ 度量指标进行科学评估，以验证增益的来源——当且仅当 $\Gamma > 1$时，候选因素才会被确认是真正的协作贡献因素，这表明改进源于真正的协作增益，而非单纯的资源累积（Malfa 等人，2025）。这种顺序验证有助于将因素归因从试错转变为更加严谨、以证据为基础的科学过程。

3. **构建因素库：** 为支持这一归因，我们倡导建立系统化的多智能体系统因素库，即一套系统组织起来、能够影响系统行为的变量集合。归因范式提供了“方法”，而这一因素库则通过系统化此前零散的设计空间，提供了“范围”，并为上述归因任务提供结构化的候选因素集。我们将因素空间划分为任务上下文（外部）和多智能体系统构建（内部）（Maldonadoet al.，2024）。后者

进一步组织为两个基本层面：控制层，代表静态架构预设（例如组织结构）（Dang et al.，2025）；信息层，描述动态执行机制（例如内容熵）（Nikitin et al.，2024；Xia etal.，2025；Kuhn et al.）。这一结构化框架为系统优化提供了严谨且可复用的路线图。

通过这一整合范式确立具体因素与协作增益之间的关联，我们旨在以系统化科学指导取代盲目试错（Panait 与 Luke，2005），推动研究者从不够系统的方法转向更主动地调控多智能体系统因素，以提升协作效能。由于迈向真正的集体人工智能科学不仅需要摆脱工程驱动的盲目试错，还要求从根本上重新思考研究方向，并就以科学为指导的多智能体系统构建达成集体共识（Zhu 等人，2026）。因此，我们倡导研究界拥抱这一整合范式，为将多智能体系统转变为一门严谨的科学提供路线图，并确保新一代集体智能建立在透明的科学逻辑之上。

2. 问题陈述

尽管多智能体系统已在众多任务中展现出卓越性能，但该领域仍严重依赖工程驱动的盲目试错，尚无法迈向严谨的科学构建（Li 等人，2024；Wu 等人，2025）。从多智能体系统构建的角度来看，这种方法遭遇了瓶颈，其根本原因在于归因歧义。这种歧义模糊了具体因素与性能增益之间的联系，严重阻碍了该领域从随机试验走向系统化科学指导的进程。具体而言，这体现为两项关键挑战：在众多因素中进行缺乏指导的搜索，以及指标无法区分真正的协作增益与单纯的资源累积（Ma 等人，2025）。

2.1. 缺乏用于因素选择的结构化分类体系

首要障碍在于因素选择的困难。由于缺乏结构化的科学指导，这一过程目前宛如盲目搜索。多智能体系统的优化受众多相互作用的系统性因素支配，例如智能体规模（Qian 等人，2025；Wang 等人，2025c）、组织结构（Dang 等人，2025；Zhou 等人，2025b）和通信机制（Zou 等人，2025；Yanget al.，2025）。因此，探索这一高维因素空间实际上已沦为盲目搜索。这种无指导的搜索会随着任务复杂度的增加而变得计算上难以处理。研究人员面临边际收益递减的风险，因为协调开销（包括通信和同步）往往比协作效益增长得更快（Yang etal.，2026）。近期研究甚至表明，当智能体规模超过特定阈值时，由于信息过载，系统性能往往会停滞甚至下降（Qian 等人，2025）。

目前，研究陷入了两难境地：我们试图通过提高系统复杂性来解决日益困难的问题，却缺乏驾驭这种复杂性的方法论（Malfa 等人，2025）。因此，我们需要一个基于因素的框架，将这种盲目搜索转变为主动且有原则的调控。

2.2. 混淆真实协作增益与资源扩展的指标

更为关键的是，除了因素选择这一挑战外，第二个缺口也阻碍了有效归因：现有指标无法区分真正的协作与资源累积。持续依赖盲目试错，主要是因为当前的评估方法往往依赖仅限于特定任务性能的指标（Sun 等人，2025；Schipper 等人，2025；Jing 等人，2025；White 等人，2025；Chan 等人，2024）。这些指标虽适用于对最终结果进行基准测试，却不可避免地混淆了两种性质截然不同的效应：协作带来的内在增益（即智能体协作所产生的能力提升），以及仅仅由计算预算增加带来的性能改善（Qian 等人，2025）。

因此，我们仍难以准确判断：多智能体系统架构是否真正通过集体智能扩展了系统的能力边界，还是仅仅相较于单智能体基线利用了更多资源——例如增加令牌消耗或智能体规模（Du 等人，2025）。这种模糊性掩盖了系统性因素与性能结果之间的因果联系，阻碍了该领域向更具预测性的科学指导迈进（Kuhn 与 Hacking，1970年）。

3. 衡量真实协作增益：指导因素归因的原则性指标

为消除这种归因歧义并超越盲目试错，我们主张构建一种诊断信号，用以量化系统真正的协作增益，并将其作为优化的严谨反馈机制（Bo 等人，2024）。尽管原始资源的累积——例如增加令牌预算或智能体数量——

确实能够提升性能，但单智能体扩展所固有的边际效用递减表明，个体能力存在一个根本上限。因此，我们提出的指标不仅是性能基准测试，更是因素归因的严谨筛选工具。通过将真正的协作增益与资源驱动的改进区分开来，该指标有助于更系统地识别促成集体智能的因素，从而为多智能体系统的设计提供更有依据的方法（Wang 等人，2025a）。

3.1. 形式化定义与理论基础

我们认为，多智能体系统研究的一个关键目标，是捕捉集体协作所促成的非线性跃迁（Qian 等人，2025）。这一探索从根本上植根于涌现理论，该理论认为，复杂系统的宏观层次行为应超越其微观层次组成部分的简单总和（Fromm，2005）。在多智能体系统的语境下，这种宏观层次行为体现为智能体间协作所带来的协作增益，而微观层次的总和则对应于各智能体单独行动时的累积表现。我们并不否认个体规模扩张与资源扩充在推动智能发展中的基础作用，但我们认为，真正的科学前沿在于识别群体层面的组织能否超越这些规模效应带来额外收益，以及收益有多大。要推动集体人工智能走向更严谨的科学研究，我们需要更好地区分这些涌现效应与单纯资源扩充所带来的贡献（Hoel 等人，2016年）。

为此，我们提出协作增益（Γ），这是一种衡量指标，用于量化在资源消耗相同的情况下，集体相较于单智能体基线的效率（Tang等人，2025）。通过将性能与这一资源等效参照进行基准测试，Γ可作为衡量真实协作增益的探针，有助于将交互带来的收益与计算规模的影响区分开来（Chen等人，2024）。

$$\Gamma = \frac{\Phi_M}{\Phi_S}^{1} \tag{1}$$

- Φ_M 表示多智能体系统的集体性能，具体体现其群体层面的组织与协作机制所带来的效能（Malfa等人，2025）。
- Φ_S 表示非协作基线性能，代表为单智能体分配与多智能体系统总资源预算相等的资源时所能达到的最佳结果。

[1]我们在此采用比率，是因为它直观简洁；但更一般化的形式可以是任意函数 $f(\Phi_M, \Phi_S)$，用于量化二者之间的关系，并可通过回归分析或其他方法学习，以捕捉更复杂的非线性交互。

在交互效应为零的零假设下，系统性能将恰好等于单智能体基线（$\Phi_m = \Phi_s$），因此有 $\Gamma = 1$ (Zhang等人，2025c）。因此，$\Gamma = 1$ 构成协作的理论下限；任何高于此阈值的值（$\Gamma > 1$）都表明系统实现了协同效应，即“整体大于部分之和”（见图2）。

为了使这一指标能够投入实际应用，我们强调，Φ. 的操作化定义和资源等价的定义都取决于任务（Zhu 等人，2025）。首先，会自适应地选择评估函数 Φ，以反映任务的内在目标（例如准确率、覆盖率或效率）（Mohammadi 等人，2025）。其次，Φ_S 的可达到最佳性能意味着采用契合任务逻辑结构的饱和基线，从而确保单智能体在预算范围内运用最有效的非协作策略（详见附录B.1）。最后，资源消耗指标——无论以令牌吞吐量、采样步数还是工具调用次数衡量——都针对各项任务的特定瓶颈进行校准（详见附录B.2）。

3.2. Γ驱动的因素归因分析

在此基础上，Γ 不仅是一项性能指标，更是用于识别集体智能实际协作收益的诊断探针。然而，仅仅计算出一个数值，并不足以支持有意义的系统设计。鉴于多智能体系统设计的复杂多变，若对未经验证的因素进行定量建模（例如回归分析），就会引入大量噪声。因此，我们建立了一个二元归因框架，根据候选因素的经验有效性，将其划分为两个不同的类别。

3.2.1. 协作增益的分类

我们将 $\Gamma = 1$ 作为协同涌现的明确边界——即架构增益与资源成本相等的等价点（Tang等人，2025）。这使我们能够将因素空间划分为两个截然不同的因素类别：

第一类：正向因素（$\Gamma > 1$）具有$\Gamma > 1$ 统计显著性的因素，代表协作增益已成功实现。此分类证实，在资源约束相同的条件下，特定设计（例如特定的通信拓扑）能够提升系统性能，使其超越单智能体的上限（Kim 等人，2025）。这些因素被视为真正的协作驱动因素，因为它们证明了自己能够推动系统突破个体能力上限。它们是可进一步进行定量优化和调整的有效候选因素。

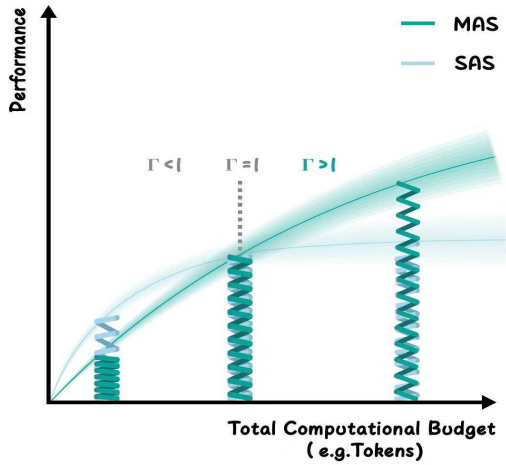


图2. 协作增益（Γ）的概念示意图。这些曲线表示多智能体系统与单智能体系统在等效计算预算下的性能，以确保可比性。当单智能体系统的性能等于或超过多智能体系统时，$\Gamma \lesssim 1$，表明仅为资源累积；反之，$\Gamma > 1$ 则表示超越单智能体的真正协作。

第二类：负向因素（$\Gamma \lesssim 1$）未能显著优于基线的因素被归入负向因素，表明归因失败，即多智能体系统的协作缺乏明显优势。这种状态以 $\Gamma \lesssim 1$为特征，代表协同红利的缺失。在这种情况下，协作的潜在益处要么被结构冗余抵消——协议仅仅重复个体推理——要么受到负向干扰的主动抑制，例如协调开销和上下文碎片化（Pan 等人，2025；Zhang 等人，2025c）。因此，负向因素构成了关键的剪枝信号，促使研究人员认识到，当前架构所增加的复杂性并无充分理由，也能避免对无效设计进行徒劳的优化。

3.2.2. 因素归因过程

为了将这一框架用于因素归因，我们主张按以下步骤开展：首先，严格将实验预设在固定的计算预算之内（例如，总令牌消耗）（Kim 等人，2025）。利用这一预算构建饱和的单智能体基线 Φ_S，确保其能够反映分配约束内最大的非协作能力，并在相同的资源分配下衡量多智能体系统的性能 Φ_M。随后，我们利用 Γ 衡量真正的协作增益，从而有效地将结构优势与单纯的资源扩张区分开来（Li 等人，2024）。为验证因素归因不受随机噪声影响，我们实施稳定性过滤阶段；如果某个因素持续保持 $\Gamma > 1$ 优势，就将其纳入正向集合。最后，这一流程以最终归因为终点，识别特定架构设计是否是经验证的因果驱动因素，能够解释观察到的协作增益。这种方法确保任何多智能体系统优化都以经验证的协作增益为依据，避免将资源浪费在对无效结构进行微调上（详见附录D）。

4. 系统化多智能体系统设计空间：结构化因素库

为促进有效的因素归因，我们建议构建一个多智能体系统因素库（见图3）。该因素库将此前碎片化的多智能体系统因素设计空间加以系统化，为后续因素归因提供一个结构化的候选因素集。

该框架使因素归因过程能够超越对盲目试错的依赖，转而借助一套严谨的分类体系促进科学探索。该分类体系包含两个主要领域：任务上下文（外部）与多智能体系统构建（内部）。在内部构建中，我们进一步将因素划分为两个基本层次：控制层，代表静态架构预设（例如组织结构）；以及信息层，刻画动态执行机制（例如内容熵）（Li等人，2024；Maldonado等人，2024）。这一分类提供了严谨的路线图，确保多智能体系统的构建由全面且合乎逻辑的设计空间所引导。关于各因素对应的具体优化策略及文献支持的实现方法，请参阅附录C。

4.1. 任务情境（外部）因素

任务属性界定了多智能体系统运行所处的外部边界条件（Decker，1995年）。这些属性塑造了问题空间的逻辑结构，并限制了协作增益的范围 Γ。具体而言，任务可分解性、顺序依赖和任务清晰度等任务属性，展现了任务上下文如何影响问题表述与协作式智能体架构之间的兼容性（Davis & Smith，1983；Malone & Crowston，1994年；Decker & Lesser，1993）。

实证证据表明，任务结构与多智能体系统设计之间的不匹配可能会抵消甚至削弱协作带来的益处（Kim 等人，2025）。例如，具有强依赖链的高度顺序化任务往往会限制并行性，降低分布式协调的有效性；而内在清晰度较低的任务通常需要明确的任务分解和规划结构（例如软件工程中的任务）（Qian 等人，2024）。

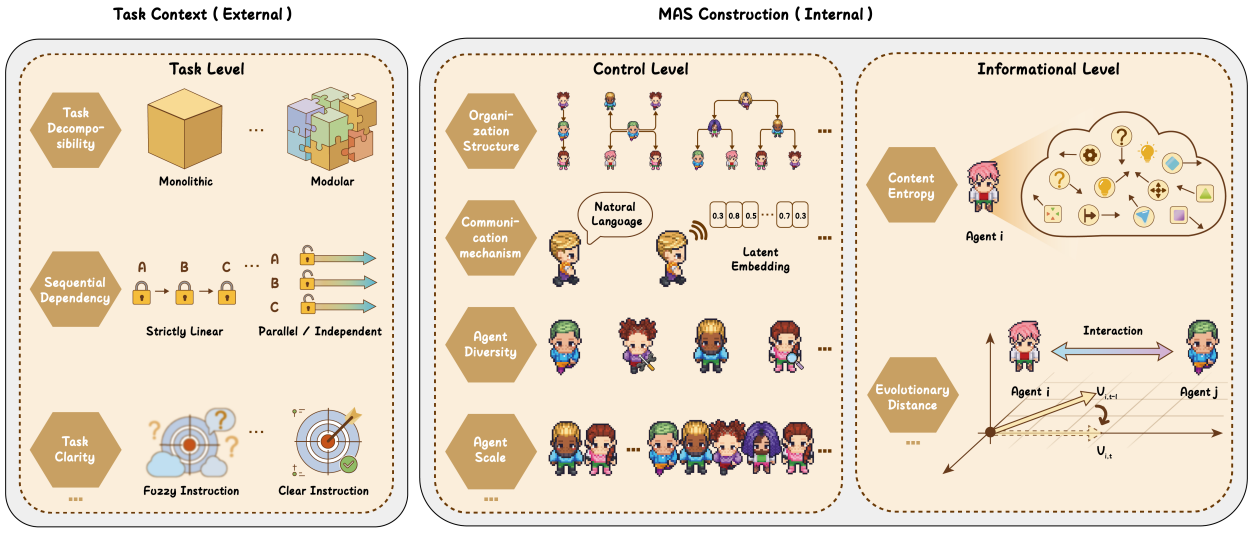


图3. 多智能体系统因素库　分类体系　. 因素分为任务上下文（外部）和多智能体系统构建（内部），其中内部维度涵盖静态控制层预设和动态信息层，指导严谨的系统设计。

这些属性似乎会塑造 Γ的反馈行为，这表明集体协作的有效性本质上受到任务执行逻辑的影响。

4.2. 多智能体系统构建（内部）因素

我们将多智能体系统因素空间系统化为两个协作层：控制层和信息层。控制层涵盖静态架构预设，作为限制协作潜力的基础（Li et al.，2024；Guoet al.，2024；Maldonado et al.，2024）。与之互补，信息层刻画动态执行机制，揭示这一潜力如何被激活并转化为实际增益（Van Dyke Parunak & Brueckner，2001；Nowak，2006）。

4.2.1. 控制层

控制层涵盖静态架构因素，作为预设协作潜力上限的基础（Li et al.，2024）。其核心逻辑将非结构化需求形式化映射为结构化系统设计，并通过特定结构因素界定操作边界。该层整合四个维度：组织结构、通信机制、智能体多样性和智能体规模（Guo et al.，2024；Maldonado et al.，2024）。设计不当会造成结构瓶颈，导致协作增益不佳。

组织结构 构成了多智能体系统的架构配置和连接拓扑，确立了智能体交互的基本路径（Li 等人，2024）。我们将其视为一个关键因素，因为即使是表现出色的智能体，其潜力也可能受到限制性拓扑结构的制约；这种结构可能造成瓶颈，限制信息传播和协作效率。（Qian 等人，2025）。

该领域的研究正在从预设的静态层级结构转向自适应的动态架构。例如，智能体超网络使系统能够以灵活的拓扑结构取代固定工作流，以仅为传统推理成本一小部分的代价实现更高的准确率（Zhang 等人，2025c）。精心构建的拓扑结构旨在优化连接密度并最大限度地减少结构冗余，从而为协作增益转化为切实的性能提升提供必要的架构基础。

通信机制支配着内部协作与环境反馈，涵盖从显式的自然语言到隐式的潜在空间表征等多种形式（Li 等人，2023）。尤其是在基于文本的交互中，不受约束的协议往往会导致“通信爆炸”：冗余对话造成资源规模呈指数级增长，而边际收益却不断递减（Kim 等人，2025；Wang 等人，2024b；Tran 等人，2025）。我们认为，过度且缺乏组织的冗余会抑制 Γ，因为关键任务信号会淹没在通信噪声中，进而使系统无法在相同预算下超越单智能体性能。

为解决这一问题，近期研究聚焦于通信高效设计，通过优化交互结构和协议，在保留协调能力的同时减少冗余（Chen 等人，2025）。例如，

对时空消息图进行单次剪枝，可筛除冗余传输，在维持系统鲁棒性的同时将令牌用量降低 28.1%–72.8%（Zhang 等人，2025b）。通过这种结构化的协调控制，该机制降低了协调成本，并旨在为将结构潜力转化为经验证的协作增益创造必要条件（$\Gamma > 1$）。

智能体多样性刻画了多智能体系统内部功能异质性的程度，是一种可调节的差异化因素。形式上，设 $\mathcal{I} = \{i_1, \ldots, i_N\}$ 为智能体集合，其中每个智能体 i_k 由元组$\mathcal{A}_k = \langle P_k$、$M_k$、$T_k$、$R_k\rangle$定义，分别表示基础参数、记忆机制、工具集和角色配置（Wang 等人，2024b；Xi 等人，2025；Schick 等人，2023）。多样性并非固定状态，而是一个可以平滑调节的范围；通过改变这些功能部分的重叠或差异程度，我们可以控制系统是保持一致，还是变得高度专业化。

我们认为，功能异质性可作为补偿个体认知偏差的机制，在拓展集体解空间的同时实现交叉验证（Jiang 等人，2023）。通过审慎调整工具集（T_k）、角色（R_k）、记忆结构（M_k）或推理模式之间的差异，可以有策略地强化智能体之间的互补优势，从而提升集体性能（Ye 等人，2025）。通过综合专业化视角，多样性为实现正向协作增益提供了一条潜在途径（$\Gamma > 1$）。

智能体规模定义了智能体总数，是交互复杂度的基础。随着规模扩大，解空间和交互路径随之扩展，为协作涌现奠定了结构基础。具体而言，协调 1,000 多个智能体可促成逻辑斯蒂增长模式，使协作涌现比传统神经网络缩放更早出现（Qian 等人，2025）。例如，OASIS 系统支持数百万个智能体，展示了更大规模如何模拟更强的集体动力学（Zhang 等人，2024）。同样，在金融交易中，TradingAgents 通过多角色协作取得了更好的表现（Xiao 等人，2025）。在这一框架中，智能体规模是一项可调参数，为通过性能相变表现为显著的集体智能开辟了路径。

4.2.2. 信息层

信息层刻画了执行过程中潜力动态转化为实际增益的过程。控制因素界定架构边界，而实时涌现的信息因素则作为系统内部认知轨迹的代理指标。我们纳入这一层，旨在提高对协作过程的机制透明度，从而定量追踪静态配置如何体现为动态协作增益

。我们分析两个维度：智能体内 维度使用内容熵衡量解的确定性和收敛情况，而智能体间 维度则使用演化距离追踪语义流和交互贡献。

内容熵通过捕捉语义特征空间中的意图分布，衡量解空间的确定性（Xia 等人，2025）。时刻 t 的瞬时内容熵定义为：

$$H_t = -\sum_i p(x_i|C_t) \log p(x_i|C_t) \tag{2}$$

其中，H_t 表示时间 t的熵，i 对智能体生成的离散内容类型集合进行索引，$p(x_i \mid C_t)$表示给定系统状态 C_t时内容 x_i 的条件概率。

在多智能体系统框架中，内容熵 H_t 可作为系统信息状态转变的定量代理指标。若 H_t 单调下降，则表明系统正沿着收敛轨迹，从随机探索走向稳定的集体状态（Navajaset al.，2018）。相反，持续的高熵意味着协调缺陷，即由于缺乏有效约束，系统处于决策分歧状态（Guo，2025）。然而，由于熵的降低仅在形式意义上描述了共识，因此并不一定等同于高质量的解决方案（Yanget al.，2025）。例如，在“上下文断裂”的情况下，智能体可能会过早忽略关键信息，从而使系统表现出伪收敛。因此，熵的简单下降无法区分有效共识与退化性信息丢失，这就需要对交互情境进行更深入的分析；附录D中提供了更详细的示例。

演化距离刻画了系统的动态“工作”，量化交互过程中的语义位移强度。信息论视角下，这一因素定义了交互行为在多大程度上会引发智能体内部状态的变化（Jaquesetal.，2019年）：

$$D_t = \sum_{i=1}^{N} \left(1 - \frac{\mathbf{v}_{i,t} \cdot \mathbf{v}_{i,t-1}}{\|\mathbf{v}_{i,t}\| \|\mathbf{v}_{i,t-1}\|}\right) \tag{3}$$

其中，D_t 表示系统在时间 t的演化距离，用于刻画相对于上一轮语义内容的整体更新强度。这里，N表示多智能体系统中的智能体总数，i 表示智能体索引。术语$\mathbf{v}_{i,t}$、$\mathbf{v}_{i,t-1}$用于衡量第 i个智能体在连续时间步的状态向量（例如语义状态嵌入）之间的余弦相似度；用 1 减去该值即可得到余弦距离，该距离反映了智能体语义状态的变化。

在理想的执行过程中，最优的演化距离意味着有意义的新信息涌现（Cover，1999）。然而，该因素的异常波动需要进行双向分析：变化过小通常表明系统陷入了冗余重复。相反，过大的变化也未必是积极信号，因为这往往意味着上下文断裂，即智能体可能失去与先前上下文的联系，导致输出与此前的逻辑完全脱节。因此，只有当演化距离与内容熵保持平衡关系时，系统才能在维持逻辑一致性的同时实现有效的性能增长（March，1991）。附录中提供了包含详细案例研究的实证分析D。

5. 其他观点

本文提出，应引导多智能体系统设计走上结构化、科学化的道路。尽管这一方法建立在坚实的逻辑基础之上，但仍可能遭到质疑。本节将回应三个主要反驳意见。

5.1. 操作复杂度与资源成本

批评者认为，我们的框架在实践中实施起来过于困难。他们指出，分析如此多种不同因素会带来大量繁重工作，而衡量具体的协作增益则需要过于严格的实验设置（Dorri 等人，2018）。从这一角度看，Γ 被视为一种不必要的负担；批评者更倾向于使用绝对准确率等更简单、更直接的指标，因为这类指标更容易追踪（Windl 等人，2022年）。

然而，我们认为，优先考虑短期便利，而不去理解系统实际发挥作用的原因，会造成严重的认知盲区。简单的指标无法揭示性能提升究竟源于单纯的资源规模扩展，还是源于真正的架构协同。与其继续盲目试错，该领域不如转而采用协作增益作为精确的反馈信号，以评估特定设计因素带来的真实收益（Fang 与 Casade- vall，2011）。尽管这种严格的评估需要开展更受控的实验，但它能提供必要的透明度，帮助区分表面改进与根本性的架构突破。这一转变对于将多智能体系统构建从经验性技艺转变为可预测、可复现的科学至关重要（Sharma 等人，2025）。

5.2. 整体论与还原论之间的矛盾

系统论的批评者认为，将多智能体系统拆解为单个因素过于简单。他们担心，若专注于各个独立部分，我们会忽视“涌现”——即系统的集体行为如何变得大于其各部分之和（Mazzocchi，2012；Pigliucci，2014）。

对此，我们认为，系统分析是理解集体力量的必要一步。正如我们通过研究个体基因来理解生命的复杂性，我们也必须找出具体的设计因素，了解它们如何相互作用并产生涌现（Fang 与 Casadevall，2011）。我们并非忽视“整体”，而是倡导一种能够提供更清晰路线图的视角，借此了解这些个体“要素”如何组合起来，形成集体智能。通过确定每个因素的作用，我们不再只是观察系统能够运作，而是进一步理解其内部组件如何促成成功（Carrodano，2025）。

5.3. 从相关性到因果发现的转变中的歧义

怀疑者认为，识别设计因素只能得出统计相关性，而非真正的因果机制，因此可能导致形成一个无法推广至新环境的静态资料库（Bai，2022年；Shu 与 Ye，2023）。

对此，我们认为，因素归因是深入理解的必要基础。我们呼吁社 Γ 区采用 Γ 作为标准反馈信号，以反映集体协作的真实成效。借助这一信号，我们可以系统地识别哪些设计因素真正推动了性能增益。这一过程为未来的因果建模提供了必不可少的数据——这种方法能够描绘具体设计“行动”与其所产生“效果”之间的直接联系。奠定这一基础将为构建更透明、更可设计的框架铺平道路，使该领域从仅仅观察系统能否运作，转向预测如何让系统运作，从而让集体智能成为可复现的成果（Sch¨olkopf，2022年）。

6. 结论

本文挑战了当前对工程驱动的盲目试错的普遍依赖，倡导通过范式转变解决多智能体系统中根本性的归因歧义。我们提出了一个综合框架，围绕三项核心战略建议展开。首先，我们主张采用协作增益指标作为科学评估工具，以区分真正的协作增益与单纯的资源累积。在此指标的基础上，我们提出一种因素归因范式，以系统地识别性能提升的真正因果驱动因素。最后，我们呼吁构建一个系统化的多智能体系统因素库，为这些归因任务提供结构化设计空间。这些建议共同旨在推动多智能体系统的构建从经验实践转向系统化科学指导。通过确保性能改进在理论上可追溯，我们为这一领域迈向严谨科学提供了路线图，促进集体智能的可靠工程化。

性能改进在理论上可追溯，我们为这一领域迈向严谨科学提供了路线图，促进集体智能的可靠工程化。

参考文献

Bai, H. 机器学习的认识论。*Filosofija.Sociologija*，2022年。

Ballerini, M., Cabibbo, N., Candelier, R., Cavagna, A., Cis- bani, E., Giardina, I., Lecomte, V., Orlandi, A., Parisi, G., Procaccini, A., 等。支配动物集体行为的相互作用取决于拓扑距离而非度量距离：一项实地研究的证据。《美国国家科学院院刊》*(PNAS)*，2008年。

Bigeard, A., Nashold, L., Krishnan, R., 和 Wu, S. 金融智能体基准：在真实世界金融研究任务上对大语言模型进行基准测试。*arXiv:2508.00828*，2025。

Bo, X., Zhang, Z., Dai, Q., Feng, X., Wang, L., Li, R., Chen, X., 和 Wen, J.-R. 基于大语言模型的反思式多智能体协作。载于神经信息处理系统大会*(NeurIPS)*，2024。

Bonabeau, E., Dorigo, M. 和 Theraulaz, G. 群体智能：从自然系统到人工系统。牛津大学出版社，1999。

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. 和 Schwaller, P. 利用化学工具增强大语言模型。2023。

Carrodano, C. 利用贝叶斯网络分析非线性因素交互的一般框架：在道路安全和公共卫生风险分析中的应用。《科学报告》，2025。

Cavagna, A., Cimarelli, A., Giardina, I., Parisi, G., Santa-gati, R., Stefanini, F. 和 Viale, M. 椋鸟群中的无标度相关性。《美国国家科学院院刊》（*PNAS*），2010。

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J. 和 Liu, Z. ChatEval：通过多智能体辩论改进基于大语言模型的评估器。发表于国际学习表征会议（*ICLR*），2024。

Chen, J., Saha, S. 和 Bansal, M. Reconcile：圆桌会议通过多样化大语言模型之间的共识提升推理能力。载于计算语言学协会年会（*ACL*），2024。

Chen, W., Su, Y., Zuo, J., Yang, C., Yuan, C., Chan, C.-M., Yu, H., Lu, Y., Hung, Y.-H., Qian, C. 等。《AgentVerse：促进多智能体协作并探索涌现行为》。载于国际学习表征会议（*ICLR*），2023。

Chen, W., Yuan, J., Qian, C., Yang, C., Liu, Z. 和 Sun, M。《Optima：优化基于大语言模型的多智能体系统的有效性与效率》。2025。Couzin, I. D., Krause, J., Franks, N. R. 和 Levin, S. A.《移动中的动物群体如何有效领导与决策》。《自然》，2005。Cover, T. M.《信息论基础》。1999。

Dang, Y., Qian, C., Luo, X., Fan, J., Xie, Z., Shi, R., Chen, W., Yang, C., Che, X., Tian, Y., Xiong, X., Han, L., Liu, Z. 和 Sun, M. 通过不断演进的编排实现多智能体协作。载于神经信息处理系统大会（*NeurIPS*），2025。

Davis, R. 和 Smith, R. G. 将谈判作为分布式问题求解的隐喻。《人工智能》，1983。

Decker, K. 和 Lesser, V. 复杂计算任务环境的定量建模。载于 *AAAI*，1993。

Decker, K. S. 以环境为中心的协调机制分析与设计。博士论文，1995年。

Di Caro, G. 和 Dorigo, M. Antnet：用于通信网络的分布式信息素控制。《人工智能研究杂志》，1998年。

Di Caro, G.、Ducatelle, F. 和 Gambardella, L. M. Anthocnet：一种受自然启发的自适应移动自组织网络路由算法。《欧洲电信汇刊》，2005。

Dorigo, M.、Maniezzo, V. 和 Colorni, A. 蚁群系统：通过协作智能体群体进行优化。《*IEEE*系统、人和控制论汇刊*B*辑》，1996年。

Dorri, A.、Kanhere, S. S. 和 Jurdak, R. 多智能体系统：综述。《*IEEE*开放获取》，2018年。

Du, Z.、Qian, C.、Liu, W.、Xie, Z.、Wang, Y.、Qiu, R.、Dang, Y.、Chen, W.、Yang, C.、Tian, Y.、Xiong, X. 和 Han, L. 通过跨团队编排实现多智能体协作。2025。

Eberhart, R. C., Shi, Y., 和 Kennedy, J. 群体智能。Elsevier，2001。

Edmondson, A. 工作团队中的心理安全与学习行为。《行政科学季刊》，1999。

Fang, F. C. 和 Casadevall, A. 分子生物学的进化。《自然综述：分子细胞生物学》，2011。

Franks, N. R. 行军蚁：一种集体智能。1989。

Fromm, J. 涌现的类型与形式。2005。

Galton, F. 民意。1907。

Ghareeb, A. E., Chang, B., Mitchener, L., Yiu, A., Szostkiewicz, C. J., Laurent, J. M., Razzak, M. T., White, A. D., Hinks, M. M., 和 Rodriques, S. G. Robin：一个用于自动化科学发现的多智能体系统。*arXiv:2505.13400*，2025。

Gong, R., Huang, Q., Ma, X., Noda, Y., Durante, Z., Zheng, Z., Terzopoulos, D., Fei-Fei, L., Gao, J., and Vo, H. Mindagent：涌现式游戏交互。载于计算语言学协会（*NAACL*）论文集，2024。

Guo, S., Shariatmadari, A. H., Xiong, G., Huang, A., Kim, M., Williams, C. M., Bekiranov, S., and Zhang, A. Ide- abench：用于研究创意生成的大语言模型基准测试。载于第*31*届*ACM SIGKDD*知识发现与数据挖掘会议论文集第*2*卷，2025。

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. 基于大语言模型的多智能体：进展与挑战综述。2024。

Guo, X. 通过条件熵缩减衡量大语言模型中的推理效用。2025。

Hoel, E. P., Albantakis, L., Marshall, W. 和 Tononi, G. 宏观层次能否胜过微观层次？跨时空尺度的整合信息。意识神经科学，2016年。

Hoff, M. P. 山地大猩猩：Karisoke 三十年的研究，2002年。

Hoff, N., Wood 和 Nagpal, R. 面向机器人群体觅食的分布式群体级算法切换。2013。

Hong, L. 和 Page, S. E. 多样化问题求解者组成的群体能够胜过高能力问题求解者组成的群体。《美国国家科学院院刊》，2004。

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z. 等。Metagpt：面向多智能体协作框架的元编程。2023。

Hu, X.-m., Zhang, J., Xiao, J., and Li, Y. 疏水-极性晶格模型中的蛋白质折叠：一种灵活的蚁群优化方法。《蛋白质与肽快报》，2008年。

Huot, F., Amplayo, R. K., Palomaki, J., Jakobovits, A. S., Clark, E., 和 Lapata, M. 智能体房间：通过多步协作生成叙事。载于第十三届国际学习表征会议（*ICLR*）。

Isaeva, V. 生物系统中的自组织。《生物学通报》，2012。Jaques, N., Lazaridou, A., Hughes, E., C¸aglar G¨ulc¸ehre, Ortega, P. A., Strouse, D., Leibo, J. Z., 和 de Freitas, N. 社会影响作为多智能体深度强化学习的内在动机。载于*ICML*，2019年。Jiang, D., Ren, X., 和 Lin, B. Y. LLM-Blender：通过成对排序和生成式融合集成大语言模型。2023。

Jing, L., Huang, Z., Wang, X., Yao, W., Yu, W., Ma, K., Zhang, H., Du, X., and Yu, D. DSBench：数据科学智能体距离成为数据科学专家还有多远？载于国际学习表征会议（*ICLR*），2025。

Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. 具有重要影响的人工智能智能体。《机器学习研究汇刊》（*TMLR*），2025。

Karaboga, D.和 ASLAN, S. 一种用于检测DNA序列中转录因子结合位点的离散人工蜂群算法。《遗传学与分子研究》，2016年。

Karaboga, D. 和Basturk, B. 一种用于数值函数优化的强大而高效的算法：人工蜂群（ABC）算法。《全球优化杂志》，2007年。

Kennedy, J. 和 Eberhart, R. 粒子群优化。载于*ICNN'95* 国际神经网络会议论文集，1995年。

Kernbach, S.、Thenius, R.、Kernbach, O. 和 Sch mickl, T. 在人工微型机器人系统中重新实现蜜蜂聚集行为。《style id='1'>《适应性行为》，2009。

Kim, Y.、Gu, K.、Park, C.、Park, C.、Schmidgall, S.、Hey- dari, A. A.、Yan, Y.、Zhang, Z.、Zhuang, Y.、Malhotra, M. 等。迈向智能体系统扩展科学。*arXiv:2512.08296*，2025。

Kuhn, L.、Gal, Y. 和 Farquhar, S. 语义不确定性：自然语言生成中用于不确定性估计的语言不变性。

Kuhn, T.S. 和 Hacking, I. 《科学革命的结构》。1970年。

Labrou, Y. 和 Finin, T. W. 智能体通信语言的语义与对话。1997。

Labrou, Y., Finin, T., and Peng, Y. 智能体通信语言：当前概况。*IEEE* 智能系统及其应用，1999。

LD, D. B. and Krishna, P. V. 受蜜蜂行为启发的云计算环境中任务的负载均衡。《应用软计算》，2013。

Li, G., Hammoud, H., Itani, H., Khizbullin, D. 和 Ghanem, B. Camel：用于探索大语言模型社会“心智”的通信智能体。神经信息处理系统大会（*NeurIPS*），2023。

Li, X., Wang, S., Zeng, S., Wu, Y. 和 Yang, Y. 基于大语言模型的多智能体系统综述：工作流、基础设施与挑战。*Vicinagearth*，2024。

Lin, B. Y., Zhou, W., Shen, M., Zhou, P., Bhagavatula, C., Choi, Y. 和 Ren, X. CommonGen：生成式常识推理的受约束文本生成挑战。在计算语言学协会（*ACL*）的 *Findings*中，2020。

Liu, F., Niu, Y., Zhang, Q., Wang, K., Dong, Z., Wong, I. N., Cheng, L., Li, T., Duan, L., Li, K. 等。面向医疗保健领域人工智能智能体的基础架构。《细胞报告·医学》，2025a。

Liu, F., Yang, Z.-R., Liu, C., SONG, T., Gao, X., 和 Liu, H. MM-agent：将 LLM 用作现实世界数学建模问题的智能体。载于神经信息处理系统大会*(NeurIPS)*，2025b。

Liu, Y. 和 Niu, B. 一种基于模拟人类社会交往行为的新型粒子群优化模型。自然与社会中的离散动力学，2012。

Lorenz, J., Rauhut, H., Schweitzer, F., 和 Helbing, D. 社会影响如何削弱群体智慧效应。美国国家科学院院刊*(PNAS)*，2011。

Lumer, E. D. 和 Faieta, B. 聚类蚂蚁种群中的多样性与适应。1994年。

Ma, M., Zhang, J., Yang, F., Kang, Y., Lin, Q., Yang, T., Rajmohan, S., 和 Zhang, D. Dover：面向 LLM 多智能体系统的干预驱动自动调试。*arXiv:2512.06749*，2025年。

Maldonado, D., Cruz, E., Torres, J. A., Cruz, P. J. 和 Benitez, S. d. P. G. 多智能体系统：关于其组成部分、框架和工作流的综述。《IEEE开放获取》，2024。

Malfa, E. L., Malfa, G.L., Marro, S., Zhang, J. M., Black, E., Luck, M., Torr, P. 和 Wooldridge, M. J. 大语言模型未能胜任多智能体任务。发表于神经信息处理系统大会*(NeurIPS)*，2025。

Malone, T.W. 与 Crowston, K. 协调的跨学科研究。*ACM*计算调查 *(CSUR)*，1994年。

March, J. G. 组织学习中的探索与利用。《组织科学》，1991年。Mazzocchi, F. 系统生物学中的复杂性与还原论–整体论之争。《*Wiley*跨学科评论：系统生物学与医学》，2012。Mech, L. D. 狼群中的阿尔法地位、支配地位与劳动分工。《加拿大动物学杂志》，1999。Minaee, S.、Mikolov, T.、Nikzad, N.、Chenaghlu, M.、Socher, R.、Amatriain, X. 与 Gao, J. 大语言模型：综述。*arXiv:2402.06196*，2024。Mirjalili, S.、Mirjalili, S. M. 与 Lewis, A. 灰狼优化器。《工程软件进展》，2014年。

Mohammadi, M., Li, Y., Lo, J. 和 Yip, W. 大语言模型智能体的评估与基准测试：一篇综述。发表于*ACM SIGKDD*知识发现与数据挖掘会议 *(SIGKDD)*，2025。

Navajas, J., Niella, T., Garbulsky, G., Bahrami, B. 和 Sig- man, M. 少数几场辩论所产生的聚合知识胜过大规模群体智慧。《自然·人类行为》，2018。

Nikitin, A., Kossen, J., Gal, Y. 和 Marttinen, P. 核语言熵：基于语义相似性对大语言模型进行细粒度不确定性量化。神经信息处理系统大会*(NeurIPS)*，2024。

Nowak, M.A. 演化动力学：探索生命的方程。哈佛大学出版社，2006。

Pan, M. Z., Cemri, M., Agrawal, L.A., Yang, S., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Ramchandran, K., Klein, D. 等。为什么多智能体系统会失败？2025。

Panait, L. 和 Luke, S. 合作式多智能体学习：研究现状。《自主智能体与多智能体系统》，2005。

Park, J. S., O’Brien, J. C., Cai, C. J., Morris, M. R., Liang, P. 和 Bernstein, M. S. 生成式智能体：人类行为的交互式拟像。2023。

Pigliucci, M. 《整体论与还原论之间：涌现的哲学入门》。《林奈学会生物学杂志》，2014。

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., and Sun, M. Chatdev：面向软件开发的通信智能体。载于计算语言学协会年会*(ACL)*，2024。

Qian, C., Xie, Z., Wang, Y., Liu, W., Zhu, K., Xia, H., Dang, Y., Du, Z., Chen, W., Yang, C., Liu, Z., and Sun, M. 扩展基于大型语言模型的多智能体协作。载于国际学习表征会议*(ICLR)*，2025。

Renney, H., Nethercott, M. N., Renney, N., and Hayes, P. 大语言模型赋能的多智能体系统：实证评估以及对新兴设计模式与范式的洞见，2026。

Reynolds, C. W. 群鸟、兽群与鱼群：一种分布式行为模型。载于 第*14*届计算机图形学与交互技术年度会议论文集，1987。

Schaller, G. B. 《塞伦盖蒂狮子：捕食者—猎物关系研究》。2009。

Schick, T., Dwivedi-Yu, J., Dess`ı, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer：语言模型可以自学使用工具。神经信息处理系统会议（*NeurIPS*），2023。

Schipper, O., Zhang, Y., Du, Y., Pechenizkiy, M., and Fang, M. Pillagerbench：在竞争性《我的世界》团队环境中对基于大语言模型的智能体进行基准测试。2025。

Sch¨olkopf, B. 机器学习的因果关系。2022年。

Seeley,T.D. 蜂群的智慧：蜜蜂群体的社会生理学。2009。

Seeley, T. D. 蜜蜂民主。2011。

Seeley, T. D., Visscher, P. K., 和 Passino, K. M. 蜜蜂蜂群中的群体决策：当10,000只蜜蜂外出寻找新居时，它们如何协作选择新的筑巢地点？《美国科学家》，2006。

Seff, A., Cera, B., Chen, D., Ng, M., Zhou, A., Nayakanti, N., Refaat, K. S., Al-Rfou, R., 和 Sapp, B. Motionlm：作为语言建模的多智能体运动预测。载于国际计算机视觉会议 *(ICCV)*，2023。

Sharma, G., Munteanu, V., Ghiasi, N. M., Banerjee, J., Varma, S., Foschini, L., Ellrott, K., Mutlu, O., Ciorb, D., 和 Ophoff, R. A. 开放科学数据库的去中心化未来。2025。

Sherman, P. W., Jarvis, J. U., and Alexander, R. D. 裸鼹鼠的生物学. PrincetonUniversity Press, 2017.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. R. 和 Yao, S. Reflexion：通过语言强化学习实现的语言智能体。载于神经信息处理系统大会*(NeurIPS)*，2023。

Shu, X. 和 Ye, Y. 知识发现：数据挖掘与机器学习方法。社会科学研究，2023。Soleimani-Pouri, M., Rezvanian, A. 和 Meybodi, M. R. 在社交网络中使用蚁群优化和粒子群优化寻找最大团。2013。Sumpter, D. J. 动物集体行为。载于动物集体行为。2010。Sun, H., Zhang, S., Niu, L., Ren, L., Xu, H., Fu, H., Zhao, F., Yuan, C. 和 Wang, X. Collab-overcooked：对大语言模型作为协作智能体进行基准测试与评估。载于自然语言处理实证方法会议（*EMNLP*），2025。

Tang, B., Liang, H., Jiang, K., and Dong, X. 评估基于大语言模型的多智能体系统时任务复杂度的重要性。载于 *NeurIPS2025* 智能体环境扩展研讨会，2025.

Tetlock, P. E. 问责对判断与选择的影响：迈向社会情境模型。载于《实验社会心理学进展》。1992.

Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B., and Nguyen, H. D. 多智能体协作机制：大语言模型综述。*arXiv:2501.06322*，2025.

Van Dyke Parunak, H. and Brueckner, S. 多智能体系统中的熵与自组织。载于第五届自主智能体国际会议论文集，2001.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., andAnandkumar, A. Voyager：基于大语言模型的开放式具身智能体。《机器学习研究汇刊》*(TMLR)*，2024a。

Wang, H., Zhao, S., Wang, J., Qiang, Z., Qin, B., andLiu, T. 超越框架：剖析多智能体系统中的协作策略，2025a。

Wang, J., WANG, J., Athiwaratkun, B., Zhang, C., and Zou, J. 智能体混合增强大语言模型的能力，2025b。

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z.-Y., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., and Wen, J. 大语言模型自主智能体综述。《计算机科学前沿》，2024b。

Wang, Q., Wang, T., Tang, Z., Li, Q., Chen, N., Liang, J., 和 He, B. Megaagent：一种无需预定义标准操作规程的大规模自主 LLM 多智能体系统。发表于*ACL*（*Findings*），2025c。

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., 和 Zhou, D. 思维链提示能够激发大语言模型的推理能力。发表于神经信息处理系统大会（*NeurIPS*），2022年。

White, I., Nottingham, K., Maniar, A., Robinson, M., Lillemark, H., Maheshwari, M., Qin, L., and Ammanabrolu, P. 一步一协作：一种用于具身推理的多智能体 LLM 框架。2025年。

Windl, M., Feger, S. S., Zijlstra, L., Schmidt, A., and Wozniak, P. W. “探索时刻并非总会到来”：设计 AI 系统的四种实用方法。发表于计算机系统中的人为因素会议*(CHI)*，2022年。

Woolley, A. W.、Chabris, C. F.、Pentland, A.、Hashmi, N. 和 Malone, T. W. 人类群体表现中存在集体智能因素的证据。科学，2010。Wu, Q.、Bansal, G.、Zhang, J.、Wu, Y.、Li, B.、Zhu, E.、Jiang, L.、Zhang, X.、Zhang, S.、Liu, J.、Awadallah, A. H.、White, R. W.、Burger, D. 和 Wang, C. AutoGen：通过多智能体对话实现新一代大语言模型应用。在语言建模会议（*COLM*），2024。Wu, Y.、Li, D.、Chen, Y.、Jiang, R.、Zou, H. P.、Huang, W.- C.、Li, Y.、Fang, L.、Wang, Z. 和 Yu, P. S. 基于大语言模型的多智能体自动驾驶系统：近期进展综述。2025。Xi, Z.、Chen, W.、Guo, X.、He, W.、Ding, Y.、Hong, B.、Zhang, M.、Wang, J.、Jin, S.、Zhou, E. 等。基于大语言模型的智能体的兴起与潜力：综述。中国科学：信息科学，2025。Xia, Z.、Xu, J.、Zhang, Y. 和 Liu, H. 大语言模型不确定性估计方法综述。2025。Xiao, Y.、Sun, E.、Luo, D. 和 Wang, W. TradingAgents：多智能体大语言模型金融交易框架。在*AAAI 2025* 第一届 *MARW*：现实世界中的多智能体人工智能研讨会，2025。Yang, X.、Li, L.、Zhou, H.、Zhu, T.、Qu, X.、Fan, Y.、Wei, Q.、Ye, R.、Kang, L.、Qin, Y.、Kou, Z.、Liu, D.、Li, Q.、Ding, N.、Chen, S. 和 Shao, J. 迈向高效智能体：记忆、工具学习与规划。2026。Yang, Y.、Ma, Y.、Feng, H.、Cheng, Y. 和 Han, Z. 减少幻觉与通信成本：基于大语言模型的多智能体中的对抗性辩论与投票机制。应用科学，2025。Yao, S.、Zhao, J.、Yu, D.、Du, N.、Shafran, I.、Narasimhan, K. R. 和 Cao, Y. ReAct：在语言模型中协同推理与行动。2022。

Ye, R.、Liu, X.、Wu, Q.、Pang, X.、Yin, Z.、Bai, L. 和 Chen, S. X-MAS：迈向使用异构大语言模型构建多智能体系统。2025。

Zhang, G., Niu, L., Fang, J., Wang, K., BAI, L.,and Wang, X. 通过智能体超网进行多智能体架构搜索。发表于国际机器学习会议*(ICML)*，2025a。

Zhang, G., Yue, Y., Li, Z., Yun, S., Wan, G., Wang, K., Cheng, D., Yu, J. X., and Chen, T. 少说废话：一种面向基于 LLM 的多智能体系统的经济型通信管线。发表于国际学习表征会议*(ICLR)*，2025b。

Zhang, H.、Cui, Z.、Chen, J.、Wang, X.、Zhang, Q.、Wang, Z.、Wu, D. 和 Hu, S. 不要过度看重多智能体辩论——我们必须重新思考评估并拥抱模型异质性。2025c。Zhang, Z.、Yin, Z. 和 Shao, J. Genss：一种通用且可扩展的基于大语言模型的智能体社交网络模拟器。载于 *NeurIPS 2024* 开放世界智能体研讨会，2024。Zhou, C.、Li, Q.、Li, C.、Yu, J.、Liu, Y.、Wang, G.、Zhang, K.、Ji, C.、Yan, Q.、He, L.、Peng, H.、Li, J.、Wu, J.、Liu, Z.、Xie, P.、Xiong, C.、Pei, J.、Yu, P. S. 和 Sun, L. 预训练基础模型综述：从 bert 到 chatgpt 的发展历程。*Int. J. Mach. Learn.Cybern.*，2025a。Zhou, H.、Wan, X.、Sun, R.、Palangi, H.、Iqbal, S.、Vuli´c、¨I.、Korhonen, A. 和 Arık, S. O. 多智能体设计：通过更优的提示词和拓扑结构优化智能体。*arXiv:2502.02533*，2025b。Zhu, C.、Hong, S.、Wu, J.、Chawla, K.、Tang, C.、Yin, Y.、Wolfe, N.、Babinsky, E. 和 Liu, D. Raffles：基于推理的 大语言模型系统故障归因，2026。Zhu, K.、Du, H.、Hong, Z.、Yang, X.、Guo, S.、Wang, Z.、Wang, Z.、Qian, C.、Tang, R.、Ji, H. 和 You, J. Multia- gentbench：评估大语言模型智能体的协作与竞争。载于 计算语言学协会年会（*ACL*），2025。Zou, J.、Yang, X.、Qiu, R.、Li, G.、Tieu, K.、Lu, P.、Shen, K.、Tong, H.、Choi, Y.、He, J. 等。多智能体系统中的潜在协作。*arXiv:2511.20639*，2025。

附录

A. 集体智能的理论背景

A.1. 生物系统中的集体智能

人工智能中的群体智能（Eberhart 等人，2001）这一概念源于对生物种群中协作性群体行为的观察（Bonabeau 等人，1999）。生物体的协作性群体行为通常可分为三种范式。

• **去中心化群体智能**。在这种范式中，系统内没有领导者或中央控制器。相反，群体行为的复杂性和功能完全源于个体之间的局部交互。集体的宏观表现远超个体能力之和。例如，行军蚁在觅食时会形成高度有序的“袭击”队形，还能动态搭建由数万只蚂蚁组成的活体桥梁，以跨越障碍物（Franks，1989）。另一个典型例子是椋鸟群飞，数千只鸟会同步、流畅且瞬时地进行空中机动，以躲避捕食者。研究表明，这种大规模协调并不需要领导者；每只鸟只需使自己的速度和方向与最近的6–7个邻居保持一致（Ballerini 等人，2008年）。这些局部规则会沿网络传播，使扰动能够传遍整个鸟群（Cavagna 等人，2010），从而产生远超单只鸟类智能的精妙集体行为。

• **中心化组织**。 在这一范式中，一名或少数几名核心个体通过其生理状态或行为，对集体的结构和功能施加决定性的自上而下影响。一个典型例子是裸鼹鼠群落，其中唯一的繁殖雌性（蚁后）主导繁殖。她会释放特定的化学信号，并表现出攻击性行为，从生理上抑制其他所有雌性的繁殖，迫使她们成为群落的工蚁（Sherman 等人，2017年）。类似地，大猩猩群体由一只被称为“银背大猩猩”的优势雄性领导。尽管他不会对个体活动进行事无巨细的管理，但对于群体何时迁移、休息或改变路线等决策，他拥有最终决定权；同时，他还会调解内部冲突，并保护群体免受外部威胁（Hoff，2002年）。

• **混合型组织**。 这一范式结合了中心化与去中心化的要素。尽管群体中有一名核心个体发挥着特殊影响力，许多重要决策仍以分布式方式集体作出。欧洲蜜蜂群落是这种混合结构的典型例子。蚁后是唯一的繁殖个体，并通过信息素维持社会凝聚力，这是中心化的标志（Seeley，2009）。然而，当群落必须寻找新的巢址时，决策过程就会变得高度民主。侦察蜂会利用象征性的摇摆舞，对候选地点进行“辩论”。最终决定并非由蚁后作出，而是在某个地点的舞蹈信号强度达到法定人数后，通过共识形成（Seeley，2011）。

尽管单个生物体通常具有有限的智能，但进化赋予了它们极其多样的通信机制，如图 4 所示。这些机制涵盖化学和物理模态，甚至包括类似于人类行为的社会学习形式，能够显式或隐式地传递丰富的信息。

在去中心化系统中，遵循简单规则的众多个体能够自组织，形成复杂且适应性强的宏观模式（Isaeva，2012），进而产生涌现智能行为。这类现象启发了种类繁多的计算算法（Sumpter，2010）。例如，自然界中的隐式通信协议通常依赖共识主动性：个体释放化学线索以改变环境，其他个体则解读这些环境变化，从而间接协调。这一机制启发了多智能体控制系统和蚁群优化算法。表 1 总结了受生物群体中观察到的合作行为启发的代表性算法。

A.2. 人类社会中的集体智能

集体智能在人类社会中同样显而易见。从宏观层次来看，数千年来的社会发展与进化本身就是集体智能的产物。文明的大多数成就并非源自某位天才，而是无数个体通过长期、大规模的协作与社会性生产逐步创造出来的。无论是在历史上还是当今实践中，群体往往胜过个体。例如，Galton 在 1907 年观察到，数百名集市游客对一头牛体重的估算平均值几乎完全准确（Galton，1907）。

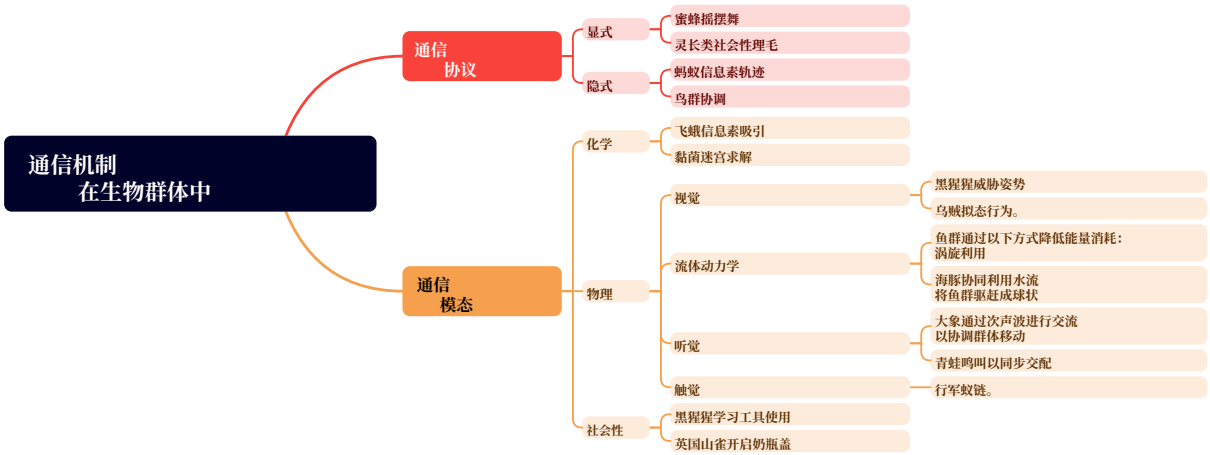


图*4.* 生物群体中通信机制的分类体系。

表1. 受生物群体行为启发的代表性算法

应用领域	代表性算法	生物启发
搜索与优化	蚁群优化（ACO）（Dorigo 等人，1995年） 粒子群优化（PSO）（Kennedy 和 Eberhart，1995年） 人工蜂群（ABC）（Karaboga & Basturk，2007年） 灰狼优化器（GWO）（Mirjalili 等人，2014）	蚂蚁通过信息素路径寻找最短路径 鸟群与鱼群的社会性搜索行为 蜜蜂群体中的智能觅食与信息共享 灰狼的社会等级与合作狩猎
机器人控制	Boids算法（Reynolds，1987） BEECLUST算法（Kernbach 等人，2009） 虚拟信息素算法（Hoff 等人，2013）	鸟群的集体飞行 温暖蜂巢区域中幼蜂的聚集行为 蚂蚁觅食中的信息素路径跟随
生物信息学	用于蛋白质折叠的蚁群优化（Hu 等人，2008年） 用于基序发现的人工蜂群算法（Karaboga 和 ASLAN，2016年） 基于蚂蚁的聚类（Lumer 和 Faieta，1994年）	蚂蚁通过信息素路径寻找最短路径 蜂群的智能觅食行为 尸体和幼虫的蚂蚁聚类
计算机网络	AntNet（Di Caro 和 Dorigo，1998年） AntHocNet(Di Caro 等人，2005) 蜜蜂行为负载均衡（LD & Krishna，2013）	蚂蚁觅食中的间接通信 基于蚁群的路径发现与维护 蜜蜂的觅食与信息共享行为
社交网络	作为社会性模型的粒子群优化（Liu& Niu，2012） 用于社区发现的蚁群优化（Soleimani-Pouri et al.，2013）	鸟类和鱼类群体中的集体运动与信息共享 蚂蚁通过信息素路径构建路径

在互联网时代，大规模在线协作使人类集体智慧的表现更加突出。网络化通信使数千人能够快速互动、分享并协作，共同创造知识、作出决策。为综合梳理人类社会中集体智能的相关研究，本节围绕五个核心维度组织文献（见图5）：

- **群体目标与任务类型。** 群体目标为集体行动指明方向，而任务类型则决定集体智能能够以何种方式——以及多容易地——展现出来。清晰且共同认可的目标能够使个人努力保持一致；模糊或存在争议的目标则会削弱专注度与协调。任务特征进一步决定适配性：需要互补的知识与技能的高度复杂任务适合群体协作；非常简单的任务，或依赖罕见个人洞察力的任务，则可能更适合个人完成。

  可分解性很重要：可分解任务受益于劳动分工与并行性；不可分解任务则需要紧密的全局协调。从组织行为与管理学的角度来看，任务复杂度与多样性为多样性、独立性和聚合创造了空间（Hong& Page，2004；Woolley 等，2010）。组织研究的证据表明，强大的团队并非由规模或层级决定，而是由心理安全感与共同意义界定（Edmondson，1999）；成员能够放心表达意见，并自下而上地推动工作。

- **群体结构。** 结构决定谁参与，以及人们如何建立联系和互动，从而影响产出和集体智能的质量。小群体往往依赖密集通信；大群体则通过层级化或网络化设计降低协调成本，其形式从严格的层级结构到松散的类市场形式不等。功能、



图*5*。 促进人类社会集体智能的机制分类体系。

认知和身份多样性能够拓展知识库、促进创新并提升集体判断。去中心化是有效开展自下而上协作的关键条件：分布式的权力和知识让个体能够依据局部信息采取行动，从而减少中心瓶颈并增强适应性。相反，当结构封闭时，权力会集中，通信受到限制；异议遭到压制，决策质量随之下降（Tetlock，1992）。

• **信息扩散与通信**。 通信机制决定知识如何在成员之间流动并实现整合。有效的通信将分散信息聚合为集体洞见，使群体能够超越个体表现。技术极大拓宽了互动的范围和深度。开放协作平台让成千上万的人能够共享文档、交流观点并实时纠正错误，在没有中央控制的情况下实现高度协调。

设计仍然至关重要：信息过载、冗余扩散和羊群效应会削弱绩效。例如，如果讨论仅围绕已知信息展开，而忽视异质信息源，集体判断就可能发生偏移（Lorenz等，2011）。良好的机制设计同时提供渠道和过滤器：它确保多样化观点进入公共空间，并通过聚合转化为群体智慧。

• **协调与决策**。 协调与决策规则决定分散的行动和意见如何转化为集体行动。协调可以是集中式的，由核心个体或制度主导，适用于紧急或高度耦合的任务；也可以是去中心化的，依靠自组织和局部反馈（Malone& Crowston，1994年）。去中心化协调常见于开放社区、预测市场和开源软件中。它强调共识、投票、市场信号或算法聚合，以将多样化意见转化为群体决策。有效的机制必须聚合异质信息，同时防止权力俘获，在广泛审议与及时收束之间取得平衡，并避免少数声音占据主导。

• **激励与规范**。 激励促使人们参与；规范确保有序且可信的合作。激励包括金钱、声誉、成就、兴趣和使命。许多开放项目依靠非货币激励，尤其是声誉，来维持大规模志愿投入，而预测市场或赏金平台则依靠金钱回报。规范通过正式规则、共同价值观和非正式惯例，使个体行为与群体目标保持一致。它们共同塑造参与深度、心理安全与集体身份，强化自下而上的协作动力。设计不当的激励与规范会加剧从众压力和表面共识，从而削弱独立性与多样性。

以下案例 s在实践中说明了这五个维度（表 le 2).

表 2. 人类集体智慧在五个维度上的比较分析。

Case	群体目标与任务类型	群体结构	信息扩散与沟通	协调与决策	激励与规范
维基百科	构建并持续更新一部开放百科全书；任务按条目模块化。	去中心化志愿者社区；一个小型数量管理员数量治理 .	公开修订历史和讨论页；透明、可追溯编辑。	共识优先，辅以投票和管理员仲裁。	内在动机和声誉；遵守政策和巡查指南。
Linux 开源社区	开发和维护操作系统内核；模块化子系统并行演进。	基于同级的分层结构；维护者负责管理子系统并集成。	公开邮件列表和代码托管记录，用于讨论和评审。	“提议、评审和合并”工作流；技术共识；维护者在需要。	内在动机与企业支持相结合；编码；编码标准和代码评审确保质量。
Waze 导航众包	实时聚合交通和地图更新；数据众包和快速迭代。	大量用户加上分级志愿编辑者共同维护地图。	应用内数据收集和社区论坛；交通信息的实时分发。	算法路线协调相结合与获许可的人工审核。	游戏化积分和声誉；编辑指南和社区规则。
预测市场	汇总概率判断事件结果的交易合约。	开放、去中心化的交易社区；无固定层级。	价格会立即反映信息，并且所有参与者都能看到。	市场出清会带来集体预测；独立交易推动快速校准。	货币激励与市场规则；反操纵保障措施和合规。
Foldit 公民科学游戏	用于寻找的游戏化搜索最优蛋白质折叠。	具有自我组织；团队并行协作。	在游戏中实时反馈；社区论坛分享策略和解决方案。	排行榜与团队合并以汇聚更优的解决方案。	徽章与排名以及科学使命；反作弊规范。

A.3. 基于大语言模型的多智能体系统中的集体智能

与依赖低维刺激—反应机制和规模效应的生物集体不同，也不同于依赖制度和文化规范的人类社会，基于大语言模型的集体智能以自然语言为协调的核心媒介，并以语言智能体——能够理解任务并进行推理的自主实体——为基本单元。此类系统通过语言共识实现内部自组织与协作，标志着智能生成进入了显式符号化和语义化的新阶段（Qian 等人，2024；Park 等人，2023；Wu 等人，2024。以下讨论将从三个方面考察其特征：交互智能体的特点、交互模式以及组织架构。

• **交互智能体的特点**。由大规模语言模型驱动的语言智能体，融合了语言理解、知识推理和任务执行等多方面能力。它们具有高度的自主性和适应性，能够在复杂的语义环境中进行目标导向推理和自我优化（Yao 等人，2022年；Shinn 等人，2023）。自主性体现在它们生成意图、独立决策和自主规划的能力上（Bran 等人，2023）。语言智能体能够独立解读以自然语言表达的任务描述、拆解目标、生成推理链，并构建执行路径的内部表征。另一方面，适应性体现在它们能够根据环境反馈和任务动态进行语言反思和策略调整，从而形成自我改进的闭环（Wang 等人，2024a；Park 等人，2023）。值得注意的是，交互智能体的认知特征通过基于语言的推理，呈现出类似人类的规划和行为调整能力。自主性与适应性的协同效应塑造了它们的智能个体性，使其能够理解任务并进行动态优化——这是基于大语言模型的集体智能不可或缺的构成要素。

• **交互智能体的交互模式**。正如人类依靠自然语言进行通信与协作，交互智能体也将自然语言作为协调的核心媒介，形成基于大语言模型的集体智能的关键机制。不同于通过固定协议（如 FIPA-ACL 或 KQML）进行通信、难以应对动态语境或任务转换的传统多智能体系统（Labrou & Finin，1997；Labrouet al.，1999），语言智能体能够在交互过程中动态生成、解释并调整语义结构，通过持续的语言协商实现语义协调（Wu et al.，2024；Li et al.，2023）。在集体层面，自然语言交互不仅是信息交换的渠道，也是推动知识共享与群体演化的力量（Park et al.，2023；Honget al.，2023）。通过持续的语义协商与反馈，交互智能体能够在去中心化条件下逐渐形成集体记忆与共享规范，进而形成稳定的协作结构。与语言在人类社会中的文化作用相似

语言智能体群体能够通过语义共识的积累，实现从个体理解到集体认知的转变。智能体之间的交互以语义显式性、动态协调和知识共建为特征，赋予集体智能更强的灵活性、适应性与社会性，形成以语言为中心的“符号集体智能”范式。

• **交互智能体的组织架构**。与生物集体类似，交互智能体群体展现出多样且灵活的组织形式。在自然界中，狼群和狮群依靠优势个体进行集中指挥与协调，是典型的集中式架构（Mech，1999；Schaller，2009）；相比之下，鸟群和鱼群通过局部感知与动态反馈形成自组织全局秩序（Couzin 等，2005；Reynolds，1987）；蜂群虽以蚁后为中心，却将符号通信机制（如信息素、摇摆舞）与分布式决策结合起来（Seeley 等，2006），展现出一种集中化与去中心化并存的混合结构。类似地，语言智能体系统的组织结构大致可分为集中式、去中心化和混合式。在集中式架构中，核心智能体充当“中央规划者”，负责统一的全局任务管理；在去中心化架构中，智能体平等互动，群体行为从个体交互中自发涌现；混合式架构则将集中式全局规划与去中心化的局部自主性结合起来，在保持任务连贯性的同时保留智能体的独立性。选择或组合最优架构，需要针对任务规模和环境复杂性，在全局控制、个体自主性与协调效率之间进行策略性权衡。

在此基础上，一个引人深思的问题随之出现：当多个基于语言模型的智能体形成一个集体时，它们的交互模式在形式与功能上几乎与人类社会的通信方式难以区分。它们同样以语言作为协调媒介，通过语义协商达成共同理解。在组织层面，它们呈现出与人类或生物集体中所观察到的类似结构分化——例如层级角色系统、去中心化协作网络以及混合动态组织形式。这种结构与交互上的相似性引出一个深刻的假说：基于语言模型的智能体集体是否也能产生一种类似于生物或人类社会中的“集体智能”?

B. 协作增益的技术细节

B.1. 任务特定基线实现

我们强调，基线 Φ_S 的操作化实例并非固定常数，而是依赖于任务的“认知上限”。它必须被自适应地选择，以反映任务的内在目标，并针对特定的资源瓶颈进行校准。基于这些原则，我们根据任务输出的性质对 Φ_S 的聚合逻辑进行分类：

• **累积型任务（例如，多目标跟踪）：** 对于性能呈线性可加性的任务，基线 Φ_S被定义为：将总资源预算分配给相互独立的单智能体实例后，各实例独立贡献之和。如果 $\Gamma > 1$，则表明协作增益（例如通过动态劳动分工避免冗余搜索路径）已成功抵消协调成本，其结果优于简单的并行处理。

• **覆盖任务（例如，数学问题集）：** 对于以任务集通过率衡量的任务，基线 Φ_S 与“等量尝试下的搜索广度”相一致。给定总预算为 N 个 token，Φ_S 通过计算单智能体在该预算下进行独立采样所解决问题的并集得出（例如，使用具有 k 条路径的自洽性）。这样可以将随机增益与“暴力采样”解耦，确保 Γ 指标衡量的是对复杂逻辑的协作性深入程度，而非单纯的概率叠加。

• **单解任务（例如代码生成、复杂规划）：** 对于需要生成单一连贯输出的任务，基线 Φ_S 代表“经过深度审议后的个体上限”。赋予单智能体与多智能体系统总预算等量的计算资源，以执行高级推理策略（例如更深入的自我反思或更长的思维链（Wei 等人，2022年））。这衡量集体智能是否通过交互实现了超越单智能体能力上限的非线性跃升，而非只是思考得更久。

B.2. 资源等价的维度

为了分离真正的 collabo 为确保增益，多智能体系统和单智能体系统基线必须校准至相同的资源消耗 n
水平。取决于 ta sk 瓶颈，“资源等价”指以下一个或多个维度：

- **模型能力对齐：**单智能体基线必须使用多智能体系统中能力最强的模型基座。这确保评估排除仅因升级底层基础模型而产生的基座模型红利或虚幻收益。

- **资源匹配评估：**单智能体必须耗尽与整个多智能体系统相等的总推理成本（例如，令牌长度、采样步数或工具调用次数）。这明确衡量资源消耗，以排除仅由资源扩张带来的冗余收益。

C. 多智能体系统因素库

本附录对第 4 节提出的因素库进行了文献综述，并介绍了具体实现。我们主张将研究范围专门聚焦于集体层面的因素，同时有意排除单智能体增强措施（如记忆或工具使用），以确保能够专注诊断集体智能。

表3. 多智能体系统优化因素分类：基于既有研究的因素库。本表总结了近期多智能体系统研究中的关键优化因素及其相关的高层策略，系统呈现了各种方法如何影响集体行为与系统性能。

类别	因素	参考文献与标题	具体优化策略
任务上下文（外部）	**任务属性**	(Liu 等,2025b) *MM-Agent*	**顺序依赖：**将开放式任务分解为四个不同的顺序阶段。
		（Bigeard 等人，2025; Kim 等人，2025）金融智能体	**高任务可分解性：** 使子任务能够执行通过集中式协调器并行运行。
		(Tang等人，2025) 深度-宽度写作 *(DW^2)*	**任务复杂度：**按深度（顺序推理步骤）和宽度（能力多样性）对任务进行建模。优先将多智能体系统用于高深度任务。
		(Huot等人) 智能体房间	**任务开放性：** 开放式任务（例如小说写作）通过专业化子任务分解。
MAS 构建（内部）	**智能体规模**	(Qian 等人，2025) *MacNet*	**通过有向无环图实现拓扑扩展：**将大规模智能体群体组织为有向无环图（有向无环图）。
		(Wang 等,2025c) *MegaAgent*	**无 SOP 动态扩展：**动态生成智能体群体以匹配实时任务复杂度。
	智能体多样性	(Ye 等, 2025) *X-MAS*	**异构模型集成：**利用模型异构性，根据领域专长为智能体角色分配专用大语言模型基于特定领域的能力。
		(Jiang 等，2023) *LLM-Blender*	**基于集成的生成式融合：**汇总多样化的模型输出，并通过成对排序和生成式融合合成出更优的回答。
		(Wang 等人，2025b) *MoA*	**分层协同优化：**统筹多样化的大语言模型构成多层结构，智能体在其中迭代优化通过关注前一层的集体输出，生成响应。
	组织结构	（Dang 等人，2025） *Puppeteer*	**强化学习驱动的动态编排：**动态安排顺序通过强化学习训练的编排器进行智能体交互，以适应不断变化的任务状态，而无需采用静态工作流。
		(Chan 等人，2024) *ChatEval*	**多智能体辩论拓扑：**将智能体组织成辩论通过.....构建综合多元观点的结构自主多轮论证。

（续见下页）

（续前页）

类别	因素	参考文献与标题	具体优化策略
		（Gong 等，2024） *MindAgent*	**集中式任务协调：**动态安排任务并下发指令，以协调多智能体与人类之间的交互。
		（Zhou 等，2025b） *MASS*	**交错式设计优化：**迭代优化局部在一个多阶段搜索过程中使用提示词、工作流拓扑和全局提示多阶段搜索过程。
		(Zhang 等人，2025a) *MaAS*	**基于超网的架构搜索：**动态采样从概率超网中采样查询依赖架构。
	通信机制	(Zhang 等人，2025b) *AgentPrune*	**单次图剪枝：**对时空消息传递图进行单次剪枝，以消除冗余交互。
		(Yang et al., 2025) 熵辩论	**熵压缩与对抗辩论：**提升通过熵压缩提升通信效率，并采用对抗辩论和投票机制进行交叉验证。
		(Chen 等，2025) *Optima*	**令牌高效训练：**利用监督微调/直接偏好优化通过成本感知奖励，在任务性能与令牌消耗之间取得平衡，以优化智能体对话。
		(Zou 等，2025) *LatentMAS*	**免训练潜在协作：**建立共享的潜在工作记忆，以传递潜在思想（最后一层隐藏嵌入）。

D. 因子归因范式的验证

本附录验证了正文提出的因子归因范式，具体展示了协作增益（Γ）作为诊断探针的效用。作为一项探索性预实验，我们系统地改变了因素库中的特定因素（组织结构、智能体多样性、智能体规模），并考察了 Γ 的相应变化，以评估将 Γ 用作控制层面因子的诊断探针的可行性。在这一特定设置下，我们的观察结果表明，该指标有助于区分真正的协作增益（$\Gamma > 1$）与资源冗余（$\Gamma < 1$）。

D.1. 验证 I：对 Γ因素的诊断敏感性

为验证 Γ 作为探测控制层面因子的有效方法的效果，我们首先进行了初步敏感性分析，并重点考察组织结构。我们在三个不同的数据集上（SRDD（Qian 等人，2024）、IdeaBench（Guo 等人，2025）和 CommonGen（Lin 等人，2020）），测量了 Γ 。



图 6。 智能体拓扑结构的视觉表示。上排展示了链式拓扑，其特点是五个不同智能体之间按顺序进行单向数据流。下排展示了树形拓扑（具体而言，是一种分支结构），其中工作流在第三个智能体之后分叉。

我们将智能体规模（5 个智能体）和模型骨干（Qwen3-30B-A3B-Instruct-2507）保持不变，仅在顺序链和分支树之间改变组织结构配置。对于每个基准测试，我们随机抽取 30 个不同的任务来评估系统性能，并报告平均指标。

在构建基线时，SAS 作为资源等效参照，并在与 MAS 相同的总令牌预算内，通过扩展思维链加以增强。这确保了 $\Gamma > 1$ 识别出的是真正的结构性收益，而非在相同资源分配下让单个智能体“思考更久”所带来的优势。

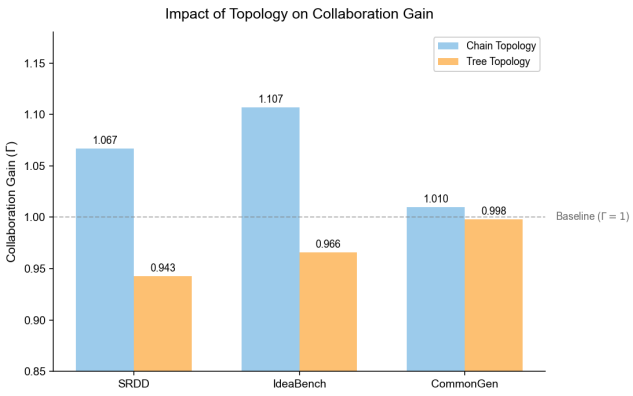


图7。 拓扑结构对协作增益的影响（Γ）

图 7 中的结果表明，该指标可能对区分所测试领域内的拓扑有效性具有一定敏感性。具体而言，链式拓扑中的 $\Gamma > 1$ 观察结果表明，这种结构有助于这些任务中的信息累积；然而，增益幅度有限，说明这种拓扑并非万能良方，在采用朴素扩展策略时仍容易崩溃（详见 D.2.4 分析）。相比之下，对于这些特定情形，该指标反映了树形拓扑（$\Gamma < 1$）中的性能惩罚，指出协调分支路径所产生的开销可能抵消并行化的潜在收益。

在这一特定实验设置中，该指标 Γ 准确反映了两种拓扑结构之间的性能差异。它捕捉到了树形拓扑中的性能下降（$\Gamma < 1$）以及链式拓扑中的性能提升（$\Gamma > 1$）。在本文评估的链式和树形拓扑背景下，这表明 Γ 有望成为一种诊断信号，用于在给定情境中识别协作效果良好或有所减弱的情况，并为后续因素分析奠定基础。

D.2. 验证II：因素归因机制：一项详细案例研究

在确认协作增益指标（Γ）具有普遍敏感性之后，我们以 SRDD数据集中的“Navigation Buddy”任务（Qian 等人，2024）为例，展开详细案例研究。我们的目标并非提出一个最先进的多智能体框架，而是将此案例用作受控环境，以说明协作增益指标如何助力因素归因，并为多智能体系统的演化式构建提供参考。我们通过逐步修改因素来分析三个阶段。

D.2.1. 实验设置与控制

实验控制与预算。 为排除计算量这一混杂因素，我们对两个系统都施加严格的令牌预算（$B_{max} = 20,\ 000$）。多智能体系统将该预算分配给所有智能体和交互轮次，而单智能体系统则获得总量相同的完整预算。至关重要的是，我们明确为单智能体系统加入思维链策略，以确保其构成“强基线”，从而确保性能差距归因于架构智能，而非计算量。

评估指标。 我们采用 ChatDev（Qian 等人，2024）中的客观指标，重点关注完整性、可执行性和一致性。我们将软件质量（Q）定义为这三个维度的乘积（Q = 完整性 ×可执行性 × 一致性）。我们使用这一综合指标计算协作增益（Γ），为评估软件可靠性及其对需求的遵循程度提供统一标准。

模型一致性。为 为确保公平，所有采用的模型均属于同一 30B参数级别。我们采用 e

Qwen3-30B-A3B-Instruct-2507 作为第一阶段中单智能体系统和多智能体系统所有角色的基础模型。

D.2.2. 第一阶段：探究“智能体多样性（角色多样性）”因素： $\Gamma > 1$

因素配置表

表 *4.* 基于第 4 节框架的多智能体系统构建（内部）

因素	配置
组织结构	链
通信机制	自然语言
智能体多样性	角色多样性（异质角色：首席执行官负责战略，首席技术官负责架构，程序员负责编码）。
智能体规模	3 个智能体（1 名首席执行官、1 名首席技术官、1 名程序员）

定性分析

我们通过检查记录的交互记录，对多智能体协作进行了定性分析。由于篇幅有限，我们仅展示部分产出，以突出多智能体系统的协作优势。每项保留的产出都附有注释，说明其战略意义。

用户 - 首席执行官：首席执行官核心产出（精选）

为 Navigation Buddy（一款实时导航系统）制定运营机制、工作流和质量约束。确保开发人员编写稳健且不走捷径的代码；保持生产可靠性和安全性。分析：集中统一战略和标准，降低实施碎片化或不一致的风险。

首席执行官 - 首席技术官：核心首席技术官输出（精选）

列表 *1.* 项目文件结构

```
{
  "文件_结构": [
    "导航助手/__初始化__.py",
    "navigation_buddy/config.py",
    "navigation_buddy/核心/__init__.py",
    "navigation_buddy/核心/position.py",
    "navigation_buddy/核心/route_planner.py",
    "导航助手/核心/流量_data.py",
    "navigation_buddy/integrations/gps_provider.py",
    "navigation_buddy/integrations/地图_provider.py",
    "导航伙伴/服务/__init__.py",
    "navigation_buddy/services/directions_service.py",
    "导航助手/服务/实时更新.py",
    "navigation_buddy/utils/__init__.py",
    "navigation_buddy/utils/日志记录.py",
    "navigation_buddy/utils/time_utils.py",
    "navigation_buddy/主程序.py"
  ]
}
```

分析：将开放式问题转化为受约束的设计空间。

首席技术官 - 程序员：核心程序员输出（精选）

列表 *2.* 输入验证与逻辑编排

```
    def get_directions(self, start, end):# 约束检查: 输入验证if
    not self.location_service.validate_location(start):raise
    值错误("起始位置无效")
```

```
        如果不是self.location_service.validate_location(end):

            引发ValueError("无效的终点位置")
        # 动态解析
        如果start == "当前位置":
            开始 = self.location_service.get_current_location()
        # 核心业务逻辑
        路线 = self.routing_engine.compute_路线(start, end)
        返回 {
            "开始": 开始 ,
            "结束": 结束 ,
            "route": 路线 ,
            "steps": self._format_steps(路线)
        }
```

分析：Thi 这体现了首席技术官定义的架构契约，降低了耦合并减少了错误传播路径。

单智能体系统（SAS）：证据输出（选定）

清单 *3.* 桩逻辑与线性逻辑

```
定义 获取_当前_位置():
    # 返回常量而非实现
    返回 "XYZ"
def 导航_to(location):
    # 忽略所需事件循环的线性脚本
    路线 = 计算_路线(current , 位置)
    打印(路线)
```

我们将生成的代码与多智能体系统的代码和资源相当的单智能体基线进行了比较。尽管采用了思维链提示，单智能体系统在长生成任务中仍出现一致性问题：随着上下文窗口逐渐填满，它会忘记规划阶段的约束，并使用占位符。

定量评估

我们以综合 Q 分数作为计算协作增益（Γ）的主要依据，该增益在第3节中提出。我们将多智能体系统与资源等效的单智能体系统进行比较。在本选定案例研究中，两个系统均能成功执行；然而，完整性和一致性方面的差异仍然存在。

表*5.* 代码生成性能的定量比较

设置	完整性	执行	缺点	Q	Γ
单智能体（基线）	0.35	1	0.76	0.27	/
多智能体系统	**0.42**	**1**	**0.81**	**0.34**	**1.26**

该多智能体系统实现了协作增益（$\Gamma > 1$），表现优于资源等效的单智能体基线。

D.2.3. 第二阶段：通过“智能体多样性（模型）”因素进行优化（$\Gamma \uparrow$）

基于第一阶段实验的Findings（其中通过角色多样性实现的功能正交性带来了正向协作增益），我们提出，智能体多样性 因素可通过模型异质性进一步优化。

因素配置更新

在第二阶段，我们将程序员智能体切换为 *Qwen3-Coder-30B-A3B-Instruct*，以引入模型异质性。此设置在保持参数规模不变的同时，单独考察专业编程专长的影响。

定性分析

这种变化并未增加表面复杂性，而是体现为涌现出同质配置中原本不存在的新工程能力类别。

新的程序员输出（选定）

```
类 实时更新:
    """管理实时导航更新"""
    定义 开始_监控(self):
        如果不是 self.is_监控:
            self.is_监控 = 真
            # 自主决策：用于非阻塞 I/O 的多线程执行
            自我监控_线程 = threading.Thread(
                target=self._monitor_循环
            )
            self.monitoring_thread.daemon = True
            self.monitoring_thread.start()
    def 监控_循环(self):
        当 self.is_monitoring:
            try:
                self._process_updates()
                time.sleep(30)
            例外 Exception 作为 e:
                打印(f"监控循环中出错：{e}")
```

分析：一种未在架构层面规定的自主并发模型。

输出展现出更高的结构完整性，线程并发和可观测性装饰器等功能便是例证。

定量评估

为严格量化增强智能体多样性因子的影响，我们重新审视基于综合软件质量得分（Γ）的协作增益（Q）指标。表6报告了对比结果。数据显示，随着智能体多样性因子逐步增强，系统性能呈阶梯式提升。

表*6*。代码生成性能的定量比较

设置	综合	执行	一致。	Q	Γ
单智能体（基线）	0.35	1	0.76	0.27	/
多智能体系统	**0.60**	**1**	**0.83**	**0.50**	**1.85**

Γ 的增加对应于模型异质性的引入。在这种情况下，这一因素似乎促进了能力从“功能正确性”向“工程抽象”的转变（例如，自主实现并发模式），表明策略性模型-角色对齐有助于提升系统性能。

D.2.4. 第三阶段：边界测试“智能体规模”因素（$\Gamma < 1$）

在验证角色异质性（第一阶段）和模型异质性（第二阶段）之后，我们进一步探讨一个更深层次的可扩展性问题：改变智能体规模因素是否能带来更高的协作增益？为找出并验证这一因素假设，我们假设扩展实现层——依次串联多个程序员智能体——应能通过迭代优化、冗余和错误纠正来提升输出质量。

因素配置更新

为隔离智能体规模的影响，实验配置更新为扩展至 5 个智能体的顺序实现链。为确保这种顺序拓扑中的协作有效，特别要求程序员 2 和程序员 3 接收前序程序员的代码，并进行增量优化（例如错误修复和逻辑强化）。

定性分析

与预期相反，定性分析表明，在这种顺序设置下，多智能体系统的性能有所下降。在这种顺序设置中，我们观察到多智能体系统的输出出现了结构性退化。在程序员 1 最初构建复杂性之后，后续智能体未能保持多文件架构。相反，最终输出被简化为单个自包含脚本，优先考虑可执行性，而非原有的结构完整性。

交互输出（选定）

```
# main.py (程序员 2)class 导航助手:def __init__
(self):self.current_location = (0.0,
0.0)self.route = ["向东行驶 2.5 公里.....","右转
进入主街。"]def main():lat_input =
float(input("纬度: "))
```

分析： 该智能体未能保留多文件架构模型。为确保可执行性，它将系统整合成一个单体脚本。

定量评估

表7. 代码生成性能的定量比较

设置	比较	执行。	考虑。	Q	Γ
单智能体（基线）	0.35	1	0.76	0.27	/
多智能体系统	**0.23**	**1**	**0.74**	**0.17**	**0.63**

协作增益骤降至 $\Gamma < 1$，这表明，在这一特定的实验设计中，简单地扩大规模并未带来累积智能。

过程动力学：基于信息流的诊断分析

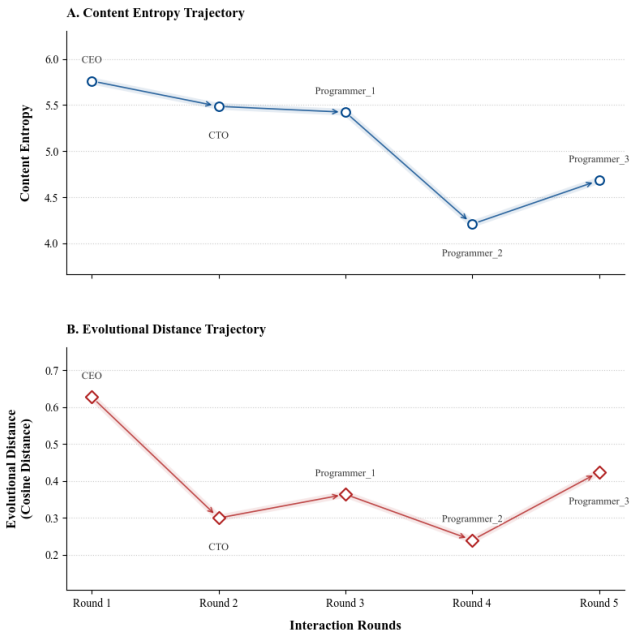


图8。信息层的演化动力学

为了解释协作增益（$\Gamma < 1$）的骤降，我们结合详细的交互日志与图8所示的信息流指标进行了诊断分析。

在初始阶段，系统呈现出稳定的传输模式：内容熵逐渐下降，而演化距离不断增加（尤其是从首席技术官到程序员 1）。这一趋势对应着健康的“语义扩展”，即抽象需求得以有效转化为具体的代码实现，同时又不

失去重点。然而，从程序员 1 到程序员 2 的过渡过程中出现了一个关键异常：内容熵和演化距离均急剧下降。结合交互日志来看，这表明上下文过载导致信息流出现了“断裂”。程序员 1 生成的庞大多文件架构似乎超出了后续智能体的有效处理窗口，导致程序员 2 舍弃了结构细节，只关注数据中范围较窄的子集。最后，程序员 3 阶段指标再次上升，表明该智能体生成了新内容，但其扩展是基于程序员 2 提供的、已经断裂且不完整的上下文。这种波动意味着，观察到的协作增益损失很可能源于一个瓶颈：难以在较长的顺序链中维持复杂上下文。

D.3. 结论

上文所呈现的分析初步表明，协作增益是多智能体系统内因素归因的一种原则性诊断信号。通过控制资源消耗，Γ 能够识别出促进或阻碍系统性能的特定因素（例如智能体多样性）。结合信息级指标，Γ 为检测内部瓶颈提供反馈，并为超越实证观察来评估与改进多智能体系统提供指导。