arXiv:2601.05606v1 [cs.MA] 9 Jan 2026

大语言模型多智能体系统中的从众动态：拓扑结构与自我–社会权重的作用

Chen Han[1,2], 谭锦[3], 俞博涵[1,4],Wenzhen Zheng[2,5],Xijin Tang[1,2][1]中国科学院大学先进跨学科科学学院[2]中国科学院数学与系统科学研究院（中国科学院）数学科学国家重点实验室[3]中国科学院大学经济与管理学院[4]中国科学院自动化研究所复杂系统认知与决策智能重点实验室[5]阶跃星辰 {hanchen23, tanjin23, yubohan23}@mails.ucas.ac.cn zhengwenzhen@stepfun.com, xjtang@iss.ac.cn

摘要

大语言模型（LLM）日益被部署为多智能体系统（MAS）中的交互式智能体，集体决策由此通过社会互动而非独立推理产生。在这一过程中，一个基础却尚未得到充分探索的机制是从众，即智能体倾向于使自身判断与当前群体意见保持一致。本文通过一项错误信息检测任务，系统研究网络拓扑如何塑造基于大语言模型的多智能体系统中的从众动态。我们提出一种置信度归一化汇聚规则，用以调节自我依赖与社会影响之间的权衡，从而能够比较两种经典决策范式：集中式聚合和分布式共识。实验结果表明，网络拓扑对集体判断的效率和鲁棒性都具有关键影响。集中式结构能够立即做出决策，但容易受到中心节点能力的影响，并表现出同模型对齐偏差。相比之下，分布式结构有助于形成更鲁棒的共识；网络连通性的提升虽能加快收敛，却也会增加错误但确信的级联风险，即智能体以较高置信度收敛至错误决策。这些发现刻画了基于大语言模型的多智能体系统中的从众动态，阐明网络拓扑与自我–社会权重如何共同影响集体决策的效率、鲁棒性及失效模式。代码将在论文录用后公开发布。

1 引言

大语言模型（LLM）日益被实例化为多智能体系统（MAS）中的交互式智能体（Wei 等人，2025年）。在众多应用中，包括协作式问题求解（Du 等人，2024年）、复杂推理（Zhang 和 Xiong，2025年）以及虚假信息检测（Li 等人，2025年），集体决策的有效性不仅取决于单个智能体的能力，也取决于其互动所引发的社会动态（Ghoshal 等人，2025年；Cisneros-Velarde，2025年）。这些动态背后的一个核心机制是从众，即智能体倾向于将自己的判断调整为接近多数意见。先前研究表明，从众可以减少个体特有的噪声并促进协同（Choi 等人，2025年）。然而，过度从众也可能引发信息级联，导致群体以高度自信收敛于错误结论（Bikhchandani 等人，2024年；Pinheiro 和 Vasconcelos，2025年）。

现有关于多智能体系统的研究主要侧重于任务效能，关注协议设计、角色专业化和协调效率（Agashe 等人，2025年；Grötschla 等人，2025年）。例如，多智能体辩论框架能够使个体判断向多数立场靠拢，在提高可靠性的同时强化系统性偏差（Han 等人，2025年）。关于说服的相关研究进一步表明，大语言模型智能体在交互过程中往往会模仿占主导地位的论证模式（Argyle 等人，2025年）。尽管计算意见动力学研究已探讨过从众现象（Helfmann 等人，2023年；Ding 等人，2025年；Han 和 Tang，2025年），现有方法仍未明确探讨智能体在决策过程中如何生成判断，尤其是交互拓扑与邻居效应如何共同影响智能体置信度的传播、聚合与放大。

在本研究中，我们通过二元虚假信息检测任务，探究网络拓扑结构如何调节从众行为，以弥补这一研究空白。我们提出一种由全局自加权参数 α 控制的置信度归一化更新规则，用以平衡智能体的自我依赖与邻域影响。具体而言，我们比较

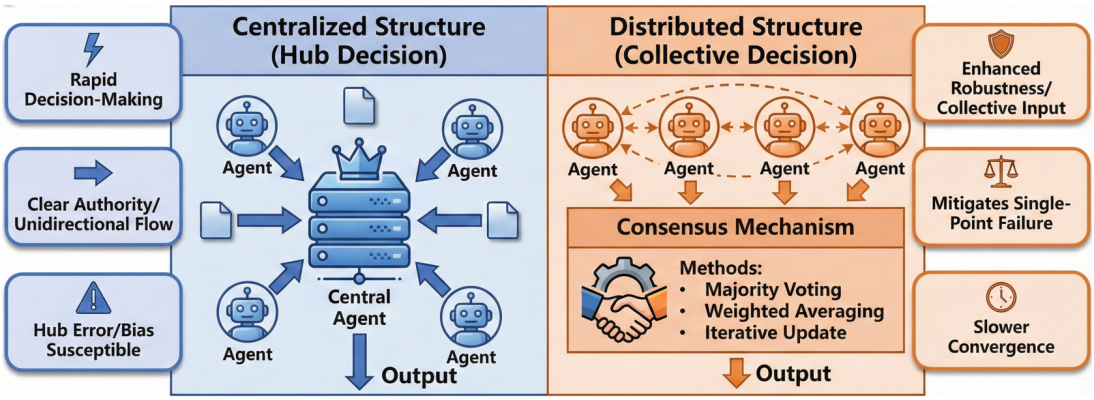

图 1：集中式（基于中心节点）与分布式（集体式）多智能体结构示意图，展示其决策机制、相对优势以及在集体推理中的固有局限性。

两种截然不同的决策模式：(i) **中心化聚合**（例如星形网络），由中心节点在单轮中综合形成集体决策；以及 (ii) **分布式共识**（例如从稀疏环形网络到完全图），影响通过迭代逐步传播。这样，我们便能区分单轮中心节点主导与多轮交互的作用。我们的贡献总结如下：

（1）拓扑结构是从众的决定因素。我们表明，网络结构通过权衡收敛速度与鲁棒性，系统性地塑造从众行为。集中式结构集中影响力，能够立即做出决策，但结果高度依赖中心节点的能力和模型同质性；分布式结构则扩散影响力，有助于形成更鲁棒的共识，而连接性的增加会加快收敛，并可能促成高置信度的错误级联。

（2）透明的置信度归一化池化规则。我们提出一种透明且可泛化的更新机制，引入一个全局参数以明确调节社会影响。该规则将智能体级置信度纳入池化过程，从而产生有界的信念分数，实现稳定的二值化，并对从众动态进行良好控制，为不同的决策范式提供了原则性依据。

（3）关于鲁棒性与风险的实证洞见。针对当前事实核查数据集开展的实验表明，系统可靠性很大程度上取决于枢纽能力和多数派构成。我们进一步刻画了从众的双重性：在可靠条件下，从众能提升准确率；但如果早期多数派信号有误，从众也可能导致自信地犯错。

2 相关研究

2.1 多智能体系统与集体决策

早期关于多智能体系统的研究主要聚焦于理想化假设下的分布式优化与共识形成，其中智能体被建模为推理能力有限的同质实体（Bao 等人，2022年；Amirkhani 和 Barshooi，2022年）。近期，大语言模型的融入拓展了这一范式：赋予智能体先进的推理能力，使多智能体系统能够应对复杂、非结构化的决策问题，例如多步问题求解（Chen 等人，2024年）和自动事实核查（Han 等人，2025年）。尽管取得了这些进展，现有研究大多强调总体性能指标，而从众等社会动态在塑造集体可靠性方面的作用仍未得到充分探讨。

2.2 从众与意见动态

从众是指个体倾向于使自己的判断与其感知到的多数意见保持一致，这一现象已在社会心理学中得到广泛研究（Capuano 和 Chekroun，2024年）。在计算领域，DeGroot 平均（Dong 等人，2024年）和有界置信度机制（Li 和 Porter，2023年）等意见动态模型，将局部交互如何产生集体共识形式化，并被用于解释信息级联和两极分化等现象（Shirzadi 等人，2025年）。然而，这些模型停留在抽象层面，无法捕捉现代大语言模型智能体所特有的语义推理或语境理解能力（Aouini 和 Loubani，2025年）。因此，经典从众理论如何拓展至基于大语言模型的多智能体系统仍有待深入探索；在这类系统中，判断是通过生成过程产生的。

2.3 网络拓扑

网络拓扑是影响影响力传播以及多智能体系统中共识形成的基本决定因素（Cheng 等人，2021年；Amirkhani 和 Barshooi，2022年）。**星形或层次图等**集中式聚合结构会集中影响力并加快决策，但也使集体结果高度依赖中心智能体的可靠性。相比之下，**分布式共识**结构（包括环形图和完全图）会将影响力分散至各个智能体，从而增强系统对局部噪声的鲁棒性，但代价是收敛速度较慢。近期研究进一步考察了连通性如何调节通信成本、收敛速度和鲁棒性（Da 等人，2025年；Wang 等人，2025年；Yang 等人，2025年）。尽管已有这些研究，现有工作仍未阐明交互拓扑、智能体置信度与从众如何相互作用，并共同塑造大语言模型驱动的多智能体系统中的集体结果。

3 方法

3.1 智能体决策模型

基于经典意见动力学和加权共识框架（Anderson 和 Ye，2019年；Li 和 Porter，2023年；Dong 等人，2024年），我们将多智能体系统（MAS）中的虚假信息检测形式化为一个二元集体决策问题。在每个交互轮次 t，智能体 i 会生成一个二元判断$y_i^{(t)} \in 0,\ 1$（其中0 表示正确，1 表示错误），并附带一个置信度分数 $p_i^{(t)}$，用于量化智能体对自身可靠性的评估。

更新规则。智能体使用置信度归一化汇聚来更新内部支持分数：

$$s_i^{(t+1)} = \frac{\alpha\, p_i^{(t)} y_i^{(t)} + (1-\alpha)\sum_{j\in N_i} p_j^{(t)} y_j^{(t)}}{\alpha\, p_i^{(t)} + (1-\alpha)\sum_{j\in N_i} p_j^{(t)} + \varepsilon}, \quad (1)$$

其中，N_i 表示智能体 i 的邻居集合，而ε 用于确保数值稳定性。该动态由两个参数控制：$\alpha \in [0,1]$ 是一个固定的超参数，用于平衡自我依赖与同伴影响；$p_i^{(t)} \in [0,1]$ 是大语言模型在每一轮生成的自我报告置信度分数，用于调节其判断的持续性以及对邻居的影响。根据定义，$s_i^{(t+1)} \in$ (0,1)，而在自我依赖程度较高（α较大）时，邻居发出的强烈且自信的信号仍可能左右决策。

二元判读。使用固定阈值 τ 将分数映射为二元标签：

$$y_i^{(t+1)} = \mathbb{1}\left[s_i^{(t+1)} \geq \tau\right], \quad (2)$$

其中，默认情况下 $\tau = 0.5$，因此 $s_i^{(t+1)} < 0.5$ 得到正确（0），而 $s_i^{(t+1)} \geq 0.5$ 得到错误（1），从而确保不同任务和置信度分布下的类别对称性以及稳定、易于解释的二值化。

3.2 提示设计

每个智能体都会收到一条结构化提示，包含三个核心部分：(i) 由大语言模型自动生成的简洁背景档案，用于将主张置于其领域背景中；(ii) 一段任务说明，指示智能体根据提供的档案和自身推理评估主张；以及 **(iii) 输出要求，指定一个二元标签** $y_i^{(t)} \in \{0,1\}$**、一个置信度分数** $p_i \in [0,1]$**和一段简短的理由**。随后，这三项输出将纳入公式（1）中的更新规则，以修正智能体的信念并指导后续决策。完整提示和伪代码见附录 A。

3.3 网络拓扑

我们分析两种不同拓扑范式下的从众动态：**集中式聚合** 和 **分布式共识**。集中式结构依赖中心节点的即时协调，而分布式架构则促进迭代式涌现共识。为确保可比性，所有拓扑均由七个固定智能体构成（$N = 7$）。评估指标的详细说明见附录 D。

3.3.1 集中式聚合

集中式聚合 的特点是信息沿单向向上流动，从外围智能体传至中央权威机构，由中央权威机构综合后决定集体结果。我们考察两种典型配置（图2）：

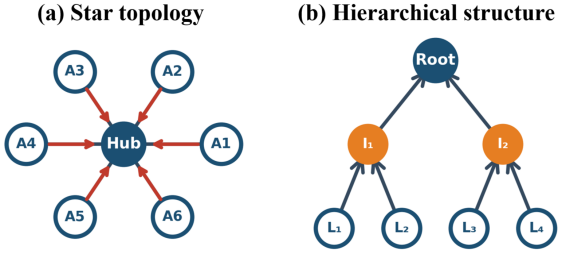

图 2：集中式聚合拓扑。

拓扑结构	α	GPT-3.5			GPT-4o			Llama3.3		
		CA	外围准确率	中心–外围一致性	中心准确率	外围准确率	中心–外围一致性	中心准确率	外围准确率	中心–外围一致性
Star	0.00	0.67	0.62	0.62	0.69	0.65	0.63	0.66	0.62	0.65
	0.25	0.73	0.68	0.73	0.75	0.71	0.77	0.74	0.72	0.68
	0.50	0.76	**0.71**	0.75	0.77	0.73	0.82	0.76	0.74	0.84
	0.75	**0.80**	0.70	**0.79**	**0.82**	**0.77**	**0.84**	**0.80**	**0.80**	**0.87**
	1.00	0.75	0.69	0.74	0.77	0.75	0.75	0.78	0.74	0.73
	无权重	0.69	0.64	0.58	0.71	0.66	0.60	0.70	0.65	0.59
层级	0.00	0.66	0.63	0.67	0.68	0.67	0.67	0.70	0.65	0.61
	0.25	0.70	0.67	0.70	0.72	0.69	0.71	0.73	0.70	0.64
	0.50	0.72	0.68	**0.77**	0.76	0.70	0.80	0.77	0.71	**0.82**
	0.75	**0.75**	**0.71**	**0.77**	**0.78**	0.73	**0.82**	**0.79**	**0.75**	0.80
	1.00	0.72	0.69	0.74	0.75	**0.74**	0.78	0.75	0.72	0.79
	无权重	0.68	0.65	0.60	0.70	0.66	0.62	0.71	0.67	0.63

表 1：单轮协议下的集中式聚合。结果以中心准确率（CA）、外围准确率（PA）和中心—外围一致性（CPC）表示。每种拓扑结构中，最佳和次佳结果分别以**粗体** 和下划线标出（并列结果也有标记）。无权重 设置移除了置信度加权聚合和全局自加权参数。在不同模型和拓扑结构中，$\alpha = 0.75$最常取得最佳性能。

(a)星形网络： 六个外围节点直接向中心枢纽传递判断，彼此之间没有横向互动。

（b）层级结构：一种三层架构，其中叶节点的输入由中间代理聚合后再传至根节点。

协议。 在这两种结构中，决策均在单个更新轮次内完成。外围或较低层级的智能体将判断提交给中心节点或根节点，由其整合这些输入并生成最终输出。该中心输出将作为群体的集体决策。

指标。 我们通过三个指标评估系统可靠性和枢纽影响力：**(1) 中心准确率（CA）：**中心节点/根节点的最终判断与真实情况相符的程度。**(2) 外围准确率（PA）：**所有非中心节点判断正确率的平均值。**(3) 中心—外围一致性（CPC）：**与中心决策保持一致的外围节点所占比例，用于量化即时从众强度。

3.3.2 分布式共识

分布式共识以对称方式分配影响力，防止任何单个节点主导决策过程。我们探讨从稀疏环形网络到完全图的连通性谱系（图 3）：

(a) 稀疏环（2 个邻居）：智能体仅与紧邻的前驱节点交互，并

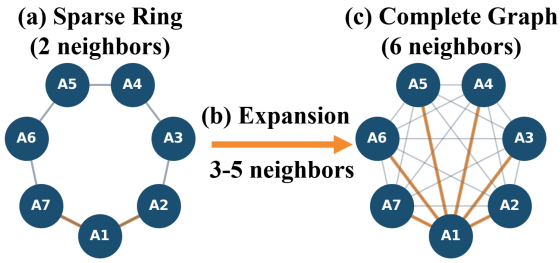

图 3：分布式共识拓扑结构。

后继节点。

(b) 扩展环（3–5 个邻居）：智能体连接到更广泛的局部邻域，从而加快信息扩散。

(c) 完全图（6 个邻居）：智能体与所有其他智能体相连，从而最大化潜在的从众强度。

协议。 决策通过迭代轮次进行。在每一步中，智能体与邻居交换状态，并根据公式（1）更新其内部置信度和判断。当达成完全共识（所有智能体持有相同的判断）或达到最大轮次限制 $T_{\max}$时，该过程终止。

指标。我们使用四项指标评估涌现共识的动态过程：**（1）最终准确率（FA）**：群体决策（ $T_{\max}$时的一致意见或多数意见）与真实情况的一致程度。**（2）达成共识时间（TTC）**：达成一致同意所需的轮次数。如果系统未能在 内收敛，则该指标未定义。

邻居	α	GPT-3.5				GPT-4o				Llama3.3			
		FA	TTC	ACI	TT	FA	TTC	ACI	TT	FA	TTC	ACI	TT
2	0.00	0.68	7.0	0.58	9.5k	0.70	6.4	0.62	9.2k	0.67	6.2	0.60	7.9k
	0.25	0.70	6.5	0.61	8.9k	0.73	6.2	0.67	9.0k	0.69	5.9	0.66	7.5k
	0.50	0.72	6.3	0.71	8.7k	0.75	5.7	0.72	8.3k	0.73	5.8	0.71	7.4k
	0.75	**0.75**	5.8	**0.75**	8.2k	**0.79**	5.1	**0.77**	7.6k	**0.78**	5.6	**0.78**	7.2k
	1.00	0.74	**5.3**	0.68	**7.5k**	0.78	**5.0**	0.75	**7.5k**	0.76	**5.1**	0.73	**6.5k**
	无权重	0.69	7.4	0.55	9.8k	0.71	6.9	0.57	9.9k	0.70	6.7	0.56	8.3k
3	0.00	0.70	7.2	0.60	10.7k	0.70	6.5	0.63	10.2k	0.68	6.3	0.61	8.8k
	0.25	0.72	6.5	0.64	9.8k	0.74	5.8	0.70	9.3k	0.70	5.9	0.69	8.0k
	0.50	0.75	5.8	0.73	9.0k	0.75	5.3	0.77	8.7k	0.76	5.6	0.74	7.8k
	0.75	**0.78**	5.1	**0.77**	8.2k	**0.82**	4.7	**0.81**	7.8k	**0.79**	5.0	**0.80**	7.2k
	1.00	0.77	**4.9**	0.74	**8.0k**	0.80	**4.5**	0.79	**7.7k**	0.78	**4.8**	0.77	**7.1k**
	无权重	0.71	6.8	0.58	10.2k	0.73	6.2	0.60	9.6k	0.72	6.1	0.59	8.4k
4	0.00	0.69	6.9	0.61	11.2k	0.71	6.3	0.63	11.0k	0.70	6.0	0.62	9.1k
	0.25	0.71	6.1	0.69	10.1k	0.74	5.4	0.73	9.6k	0.72	5.7	0.73	8.4k
	0.50	0.74	5.0	0.74	8.8k	0.81	4.7	0.77	8.6k	0.79	5.0	0.76	7.5k
	0.75	**0.78**	4.3	**0.79**	7.9k	**0.83**	4.1	**0.81**	7.9k	**0.81**	4.2	**0.82**	6.9k
	1.00	0.77	**4.0**	0.75	**7.4k**	0.82	**3.9**	0.80	**7.6k**	0.80	**4.1**	0.79	**6.8k**
	无权重	0.72	6.2	0.60	10.2k	0.74	5.7	0.62	9.9k	0.73	5.6	0.61	8.4k
5	0.00	0.69	6.8	0.59	11.7k	0.71	6.2	0.62	11.4k	0.68	6.1	0.60	9.8k
	0.25	0.72	5.5	0.70	10.0k	0.78	5.3	0.73	10.3k	0.74	5.1	0.74	7.9k
	0.50	0.74	4.5	0.76	8.6k	0.81	4.1	0.79	8.8k	0.82	4.4	0.78	7.6k
	0.75	**0.75**	3.7	**0.81**	7.6k	**0.84**	3.5	**0.83**	7.8k	**0.84**	3.4	**0.84**	6.6k
	1.00	0.74	**3.2**	0.78	**6.9k**	0.83	**3.0**	0.81	**7.0k**	0.81	**3.1**	0.80	**6.2k**
	无权重	0.73	5.8	0.63	10.4k	0.75	5.2	0.65	10.3k	0.74	5.3	0.64	8.5k
6	0.00	0.72	6.3	0.62	11.4k	0.74	5.8	0.65	11.1k	0.71	5.7	0.63	9.0k
	0.25	0.75	5.0	0.74	9.4k	0.77	4.6	0.77	9.3k	0.78	4.7	0.78	7.3k
	0.50	0.75	4.0	0.80	7.8k	0.82	3.8	0.77	8.0k	0.81	4.2	0.81	6.7k
	0.75	**0.78**	3.8	**0.81**	7.7k	**0.85**	3.4	**0.85**	7.2k	**0.83**	3.2	**0.82**	6.1k
	1.00	0.77	**3.2**	0.78	**6.7k**	0.80	**2.9**	0.83	**6.8k**	0.82	**3.0**	0.81	**5.6k**
	无权重	0.74	5.4	0.66	10.1k	0.76	5.0	0.68	9.8k	0.75	5.1	0.67	7.8k

表 2：不同邻居数量 m 和自加权设置下的分布式共识结果。性能通过最终准确率（FA）、达成共识时间（TTC）、平均从众指数（ACI）和总令牌数（TT）进行报告。对于每个邻居数量区块，最佳和次佳结果分别以**粗体** 和下划线标出（并列结果也会标记）。无权重 设置取消了置信度加权聚合和全局自加权参数。在不同模型和拓扑结构中，$\alpha = 0.75$ 最常取得最佳性能。

（3）一致性指数（CI）：第 t 轮的群体内部一致程度，定义为采纳多数标签的智能体所占比例。由于判断为二元，CI 的取值范围为 $[0.5, 1]$，其中 CI $= 1$ 表示一致同意，而 CI $= 0.5$ 对应群体意见各半。我们报告**平均一致性指数（ACI）**，即截至 $T_{\max}$ 的各轮 CI 均值。**（4）总令牌数（TT）**：累计成本，即所有智能体输出和交互轮次中的令牌使用量总和。

4 实验

4.1 实验设置

数据集。 我们收集了 Snopes25，这是一个新基准测试集，包含 448 条由 Snopes 编辑核查过的真实世界主张（252 条错误，196 条正确）。所有主张均来自 2025年1月至6月，以尽量减少预训练知识可能造成的数据污染。

实现。 实验采用两个专有模型 GPT-3.5（OpenAI，2023年）和 GPT-4o（OpenAI，2024年），以及一个开源模型 Llama3.3-70B-Instruct（AI@Meta，2024年），以确保所有模型的训练截止时间均早于我们的评估数据。详细的超参数和零样本基线见附录 B。对于**集中式聚合**，决策仅限于单次更新轮次，以捕捉对中心节点或根节点的即时从众。对于**分布式共识**，智能体会迭代交换判断，最多进行 $T_{\max} = 10$ 轮次，或直至达成共识。

参数配置。 我们考虑 α 的五个取值：0（完全从众）、0.25（受同伴影响）、0.50（平衡）、0.75（自立）和 1（完全独立），以体现对同伴影响与自我依赖的不同倾向。所有报告的指标均为 10 次独立运行的均值，

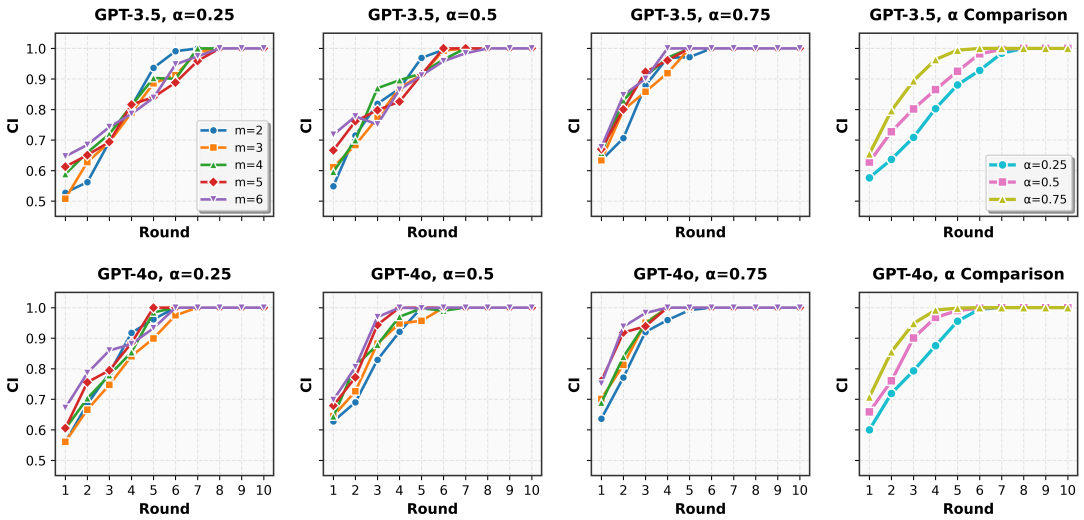

图4：中　心化的从众指数在不同网络密度下随迭代轮次的变化　　nd
自我—社会 加权，揭示了早期快速对齐以及密集结构中边际收益递减的现象。

标准误均为 < 0.09。为单独考察置信度加权机制的作用，我们引入一个无权重 基线，同时移除自我加权参数 α 和置信度分数 p，使智能体更新完全取决于邻居交互的文本语境，从而单独考察显式置信度加权的效果。

4.2 集中式聚合结果

我们首先评估**集中式聚合**，结果汇总于表1。在两种拓扑结构和所有模型中，中心准确率（CA）均随自我加权参数 α的增大而单调上升，这表明更强的自我依赖使中心节点能够更好地过滤外围噪声。例如，在星型拓扑中，GPT-3.5 的 CA 从 $\alpha = 0$ 时的 0.67 提高到 $\alpha = 0.75$时的 0.80。层级结构也呈现相同趋势，但由于信息在不同层级间稀释，提升幅度有所减弱。

相比之下，外围准确率（PA）在 α 中总体保持稳定，这表明自我–社会权重的变化主要影响中心节点汇总结果的质量，而非提升局部智能体的正确性。中心—外围一致性（CPC）随 α 增大而在不同模型和拓扑结构中上升，这表明随着中心节点更加重视自身基于置信度加权的判断，其最终决策会更加稳定；如果中心节点具备足够能力，也更有可能与外围投票保持一致。

4.3 分布式共识的结果

接下来，我们转向**分布式共识**，表2 总结了不同网络密度和自我—社会权重下的性能。最终准确率（FA）通常会随连通性 m 提高，并且往往在自我依赖程度中等偏高时达到峰值（$\alpha = 0.75$）。例如，在受社会影响的设置（$\alpha = 0.25$）下使用 GPT-3.5 时，连通性从稀疏环形网络（$m = 2$）提升至完全图（$m = 6$），准确率从 0.70 提高到 0.75。更稠密的网络有助于信息快速传播，从而加快收敛；而降低 α 往往会加强从众压力，进一步加快共识形成，但也可能放大早期偏差。在稀疏网络（$m = 2$）中，若同伴影响力较强（$\alpha = 0.25$），从众程度仍较为适中；而在高度连通的网络中（$m = 6$），群体则表现出较高的同质性。

没有置信度加权机制时，FA 在所有拓扑结构下都会一致性下降，而单纯增加连通性并不能恢复完整模型带来的收益。虽然增加邻居会提高每轮消息量，但更高的m 会显著减少达成共识所需的轮次。总而言之，稀疏网络能够保留多样性并抑制过早收敛，而完全图可以加快共识形成，但在早期信号存在偏差时，可能更容易受到信息级联的影响。

图4 进一步展示了在不同邻居数量下，从众指数（CI）随轮次变化的时间演化 m 以及自我—社会

设置	模型分配			性能				
	Hub	Left	右侧	CA	PA_L	PA_R	CPC_L	CPC_R
Cap-H1	GPT-4o	GPT-4o	GPT-3.5	0.82	0.79	0.67	0.87	0.74
Cap-H2	GPT-3.5	GPT-3.5	GPT-4o	0.77	0.73	0.77	0.81	0.69
Cap-H3	Llama3.3-70B	Llama3.3-70B	Llama3.3-8B	0.80	0.76	0.70	0.86	0.78
Cap-H4	Llama3.3-8B	Llama3.3-8B	Llama3.3-70B	0.76	0.71	0.75	0.80	0.72
Type-H1	GPT-4o	GPT-4o	Llama3.3-70B	0.82	0.75	0.73	0.88	0.76
Type-H2	Llama3.3-70B	Llama3.3-70B	GPT-4o	0.79	0.72	0.76	0.84	0.77

表 3：异构集中式聚合。左侧分支与中心节点保持一致（骨干模型相同），右侧分支则分配了不同的模型。我们报告中心准确率（CA）、分支层面的外围准确率（PA_L、PA_R）以及中心—外围一致性（CPC_L、CPC_R）。集中式性能主要由枢纽能力驱动，而非外围模型的实力。

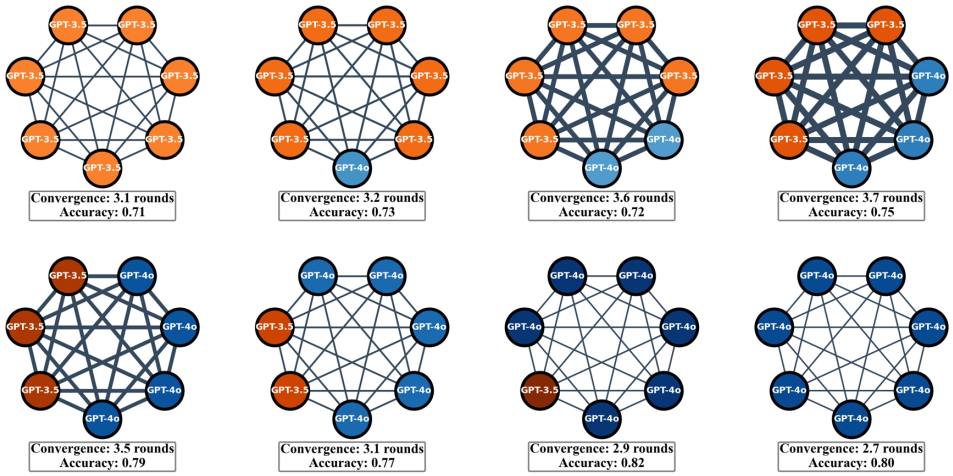

图 5：异构模型构成下的分布式共识。节点颜色越深，表示准确率越高；边越粗，则表示达成共识时间越长。较强智能体的比例和结构均会共同影响收敛速度和最终集体准确率。

加权。网络越密集，CI 上升越快，这表明一旦出现主导意见，群体一致性便能更快巩固。具体而言，稀疏环形网络（$m = 2$）通常从 0.55–0.65 起步，可能需要六轮或更多轮次才能接近一致同意；而更密集的结构（$m \geq 4$）在第二轮时便超过 0.75，并在第四轮时几乎收敛。在所有设置中，CI 增幅最大的是前四轮，之后进入较慢的巩固阶段，直至达成共识；当网络密度足够高时，提高 α或扩大 m带来的收益会逐渐递减。

5 讨论：影响多智能体系统从众的因素

5.1 模型异质性

为刻画基于大语言模型的多智能体系统中的模型异质性，我们使用不同的大语言模型骨干模型来实例化智能体。在**集中式聚合**中，中心节点

与左侧分支（即对齐分支）共享同一骨干模型，而右侧分支则使用不同的模型进行实例化（即对立分支）。我们将自我–社会权重固定为 $\alpha = 0.75$，并针对每项主张，报告中心节点层面的中心准确率（CA）、分支层面的外围准确率（PA_L、PA_R）和中心–外围一致性（CPC_L、CPC_R）；每项主张均进行 10 次独立运行。

表3揭示了与集中式从众相关的两个显著现象。首先，出现了一种稳定的同模型对齐偏差：中心节点与对齐分支保持一致的频率高于与对立分支保持一致的频率，$_L$ 在所有设置中，中心–外围一致性均持续高于中心–外围一致性 $_R$。这一模式表明，中心节点更容易接受与自身推理风格和归纳偏差相似的同伴推理依据（Laurito 等人，2025年）。其次，系统可靠性主要受中心节点能力的限制：即使外围包含一个更强的

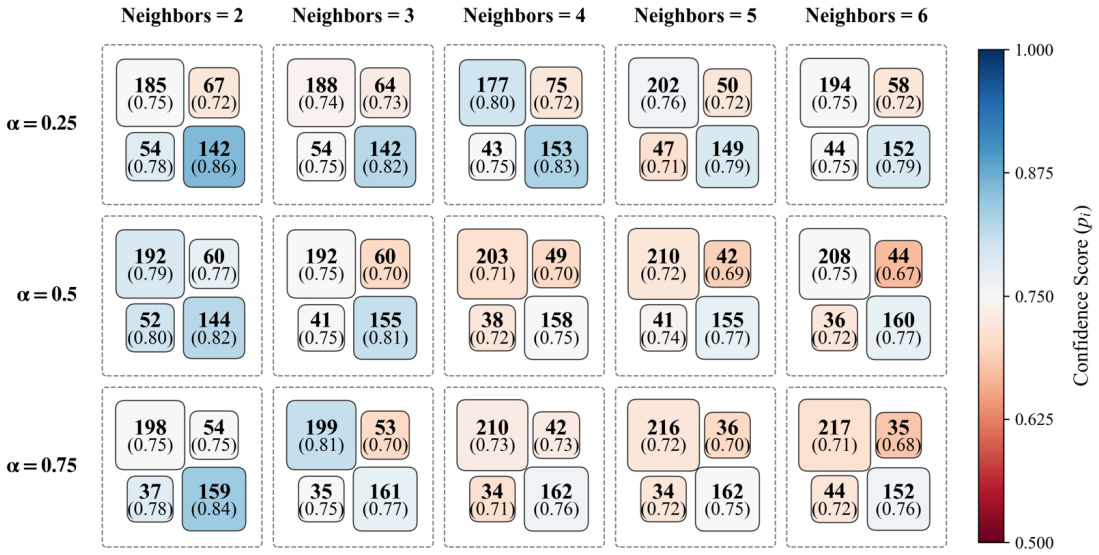

图 6：GPT-4o 在**分布式共识**下，随邻居数量和 α值变化的混淆矩阵热图。每个单元格展示群体层面的结果：正确分类的真实主张数量（左上）、正确分类的虚假主张数量（右上）、错误分类的真实主张数量（左下），以及错误分类的虚假主张数量（右下）。单元格大小表示出现频率，颜色则表示平均置信度分数 p_i。更高的连通性和更低的自我依赖会导致更多高置信度误分类，凸显了从众的一个关键失效模式。

模型并位于对立分支上，中心节点采用更强模型时，中心准确率仍然更高。总之，这些结果凸显了集中式多智能体系统的一个弱点：外围多样性无法可靠地弥补中心节点能力不足；相反，中心节点可能充当选择性过滤器，放大一致输入，同时降低不同观点的权重。

我们进一步考察分布式共识中的异质性，在完全图中改变 GPT-3.5 与 GPT-4o 的比例（0:7 至7:0）。实验针对 50 条主张开展，每条主张运行 10 次，取值为 $\alpha = 0.75$。如图5所示，异质群体的收敛速度比同质群体更慢，这与不同的推理先验和置信度校准所带来的审议摩擦加剧相一致。相比之下，同质群体收敛得更快，而 GPT-4o 智能体占多数的群体始终能达到更高的最终准确率。这表明，尽管同质性主要决定收敛速度，但涌现共识的准确率最终取决于占主导地位的模型类别的能力。

5.2 从众的双刃剑效应

当偏差信号占据主导时，从众可能会将个体错误放大为系统性的集体失效（Han 等人，2025年）。我们通过混淆矩阵热图（图 6）考察了 GPT-4o 在**分布式共识**下的这些失效模式。首先，连通性增强使智能体接触到更广泛的同伴信号，从而削弱个体证据，同时促进群体层面的协调。一旦接触到更广泛的同伴信号，这会削弱个体证据，同时促进群体层面的协调。其次，更大的 α（更强的自我依赖）通常能提高准确率，这证实了保持足够的独立性对于抵御误导性的社会影响至关重要。然而，群体有时会以高度确定性收敛于虚假主张，使错误信息根深蒂固，而不是加以纠正。这些模式揭示了从众的矛盾性质：在条件平衡时，它能提升可靠性；但当早期偏差占据主导时，也可能固化集体幻觉。

6 结论

我们的研究表明，基于大语言模型的多智能体系统中的从众行为由网络拓扑和自我–社会权重共同塑造。在集中式聚合中，集体结果的可靠性与中心节点的能力密切相关，能力更强的中心节点会提升整体系统准确率。在分布式共识中，从众行为源于迭代式局部交互；更高的连通性既能提高收敛效率和平均准确率，也会加剧系统对信息级联的易感性。尽管从众有助于达成共识，但当初始偏差占据主导时，也可能导致高置信度错误被固化。因此，实际的多智能体系统设计必须通过谨慎的拓扑设计、权重校准和对过早收敛的控制，在效率与鲁棒性之间取得平衡。

局限性

建模假设。 首先，所提出的置信度归一化汇聚规则依赖于大语言模型生成的自我报告置信度，但其校准准确性无法得到保证，且可能部分反映的是表达风格上的自信，而非对正确概率的忠实估计。此外，自我—社会权重参数在每次运行中固定，并由所有智能体共享；大多数分布式共识实验将交互限制在固定的结构化逐轮更新模板中，而非开放式审议、提问或借助工具搜寻证据。

规模与模型选择。 其次，所有实验均采用固定的群体规模（$N = 7$），且仅使用三种大语言模型。此设计基于三方面考虑。首先，受协同和成本限制，当代许多多智能体系统的部署规模仍然不大。其次，固定 $N = 7$ 能够严格遵循三级层级配置（1 个根节点、2 个中间聚合器和 4 个叶节点）。第三，选择这些模型主要是为了避免因近期知识截止日期而产生知识泄漏。我们的目标并非最大化检测准确率，而是考察智能体层级交互机制如何塑造集体行为。因此，尽管更强大的模型可能会改变整体性能水平，我们预计本研究发现的相对交互效应在定性层面上仍然稳健。

任务与领域范围。 第三，本研究仅限于二元虚假信息检测。尽管这项任务非常适合用于分离从众动态，但它仅代表一类狭窄的集体推理问题。多跳推理、政策分析或创造性协作等更开放式的任务，可能呈现出定性上不同的交互模式；在这些模式中，从众会与探索、角色专业化或策略行为相互作用，而这些现象超出了当前框架的范围。

参考文献

Saaket Agashe、Yue Fan、Anthony Reyna 和 Xin Eric Wang。2025年。LLM-coordination：评估和分析大语言模型中的多智能体协调能力。载于计算语言学协会年会北美分会 *2025* 年论文集，第 8038–8057 页，阿尔伯克基，新墨西哥州。计算语言学协会。

AI@Meta。2024年。Llama 3 模型卡。

Abdollah Amirkhani 和Amir Hossein Barshooi。2022年。多智能体系统中的共识：综述。人工智能评论，55(5):3897–3935。

Brian DO Anderson 和 Mengbin Ye。2019年。影响力网络上意见动态建模与分析的最新进展。《自动化与计算国际期刊》，16(2)：129–149。

Mourad Aouini 和 Jinan Loubani。2025年。迈向更具情境性的智能体：一种提取器—生成器优化框架。预印本，arXiv:2502.12926。

Lisa P. Argyle、Ethan C. Busby、Joshua R. Gubler、Alex Lyman、Justin Olcott、Jackson Pond 和 David Wingate。2025年。使用人工智能检验政治说服理论。《美国国家科学院院刊》，122(18):e2412815122。Guangyan Bao、Lifeng Ma 和 Xiaojian Yi。2022年。受约束异构多智能体系统协同控制的最新进展：综述。《系统科学与控制工程》，10(1):539–551。Sushil Bikhchandani、David Hirshleifer、Omer Tamuz 和 Ivo Welch。2024年。信息级联与社会学习。《经济文献杂志》，62(3):1040–93。Carla Capuano 和 Peggy Chekroun。2024年。从众研究的系统综述。《社会心理学国际评论》，37(1)。Pei Chen、Shuai Zhang 和 Boran Han。2024年。CoMM：面向复杂问题求解的协作式多智能体、多推理路径提示。载于计算语言学协会年会北美分会 *2024* 年论文集，第 1720–1738 页，墨西哥城，墨西哥。计算语言学协会。Li Cheng、Yijie Wang 和 Xinwang Liu。2021年。用于无监督多视图异常检测的邻域共识网络。*AAAI* 人工智能会议论文集，35(8):7099–7106。Min Choi、Keonwoo Kim、Sungwon Chae 和 Sangyeop Baek。2025年。多智能体系统中群体从众的实证研究。载于计算语言学协会年会 *2025* 年论文集，第 5123–5139 页，维也纳，奥地利。计算语言学协会。Pedro Cisneros-Velarde。2025年。大语言模型多智能体系统意见动态中的偏差：资金分配案例研究。载于计算语言学协会年会北美分会 *2025* 年论文集，第 1889–1916 页，新墨西哥州阿尔伯克基。计算语言学协会。Longchao Da、Xiaoou Liu、Jiaxin Dai、Lu Cheng、Yaqing Wang 和 Hua Wei。2025年。理解大语言模型解释的不确定性：基于推理拓扑结构的视角。预印本，arXiv:2502.17026。

Guozhu Ding、Zuer Liu、Shan Li、Jie Cao 和 Zhuohai Ye。2025年。心态类型与社群构成对意见动态的影响：一项基于大语言模型的多智能体模拟研究。《人类行为中的计算机》，172:108730。

Yucheng Dong、Zhaogang Ding 和 GangKou。2024年。《社会网络DeGroot模型》。*Springer*图书。

Yilun Du、Shuang Li、Antonio Torralba、Joshua B. Tenenbaum 和 Igor Mordatch。2024年。通过多智能体辩论提升语言模型的事实性和推理能力。载于第*41*届国际机器学习会议论文集，ICML'24。JMLR.org。

Arnab Kumar Ghoshal、Nabanita Das、Soham Das 和 Subhankar Dhar。2025年。基于网络拓扑结构的社交网络错误信息传播最小化方法。社交网络分析与挖掘，15(1):15。

Florian Grötschla、LuisMüller、Jan Tönshoff、Mikhail Galkin 和 Bryan Perozzi。2025年。Agentsnet：多智能体大语言模型中的协调与协作推理。预印本，arXiv:2507.08616。

Chen Han 和 Xijin Tang。2025年。基于大语言模型的错误信息易感性测试智能体模拟框架：心理因素带来的启示。载于《知识与系统科学》，第 355–368 页，新加坡。Springer Nature Singapore。

Chen Han、Wenzhen Zheng 和 Xijin Tang。2025年。Debate-to-detect：将错误信息检测重新表述为基于大语言模型的现实世界辩论。载于*2025*年自然语言处理实证方法会议论文集，第 15125–15140 页，苏州，中国。计算语言学协会。

Luzie Helfmann、Nataša Djurdjevac Conrad、Philipp Lorenz-Spreen 和 Christof Schütte。2023年。影响者和媒体策略作用下的意见动态建模。*Scientific Reports*，13:19375。

Walter Laurito、Benjamin Davis、Peli Grietzer、Tomáš Gavenˇciak、Ada Böhm 和 J an Kulveit。2025年。人工智能间偏见：大语言模型偏好由大语言模型生成的通信内容。美国国家科学院院刊，122(31):e2415697122。

Grace J. Li 和 Mason A. Porter。2023年。具有异质节点活跃度的意见动态有界置信度模型。《物理评论研究》，5：023179。

Hui Li、Ante Wang、kunquan li、Zhihao Wang、Liang Zhang、Delai Qiu、Qingsong Liu 和 Jinsong Su。2025年。一种用于跨领域虚假信息检测的多智能体框架，采用自动化决策规则优化。预印本，arXiv:2503.23329。

OpenAI。2023年。Gpt-3.5 turbo。https://platform.openai.com/docs/models/gpt-3-5-turbo。访问日期：2025年9月10日。

OpenAI。2024年。Hello gpt-4o。https://openai.com/index/hello-gpt-4o/。访问日期：2025年9月10日。

Flávio L. Pinheiro 和 Vítor V. Vasconcelos。2025年。异质更新过程塑造社交网络中的信息级联。*Scientific Reports*，15(1)：13999。

Mohammad Shirzadi、Emilio Cruciani 和 Ahad N. Zehmakan。2025年。意见动态：全面综述。预印本，arXiv:2511.00401。

Shilong Wang、Guibin Zhang、Miao Yu、Guancheng Wan、Fanci Meng、Chongye Guo、Kun Wang 和 Yang Wang。2025年。G-safeguard：一种面向基于大语言模型的多智能体系统的拓扑引导安全性视角与治理方法。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），第7261–7276页，奥地利维也纳。计算语言学协会。

Xinghai Wei、Tingting Yuan、Jie Yuan、Dongxiao Liu 和 Xiaoming Fu。2025年。Rescom：用于高效多智能体通信学习的奖励塑造课程。在《第 24 届自主智能体与多智能体系统国际会议论文集》中，AAMAS '25，第 2153–2161 页，南卡罗来纳州里奇兰。自主智能体与多智能体系统国际基金会。

Jiaxi Yang、Mengqi Zhang、Yiqiao Jin、Hao Chen、Qingsong Wen、Lu Lin、Yi He、Weijie Xu、James Evans 和 Jindong Wang。2025年。拓扑结构学习应成为基于大语言模型的多智能体系统的研究重点。预印本，arXiv:2505.22467。

Shaowei Zhang 和 Deyi Xiong。2025年。Debate4MATH：面向数学细粒度推理的多智能体辩论。收录于计算语言学协会年会 *2025* 年论文集，第 16810–16824 页，奥地利维也纳。计算语言学协会。

A 提示词模板

A.1 集中式聚合

A.1.1 外围智能体（叶节点）：单轮判断

用途：叶节点/外围智能体给出单轮决策；不会看到任何同行输入。

提示词。

你是一名谨慎的事实核查代理。请阅读主张，并给出二元↪判断和校准置信度。

[任务]1) 判断该主张是正确（=0）还是错误（=1）。↪

2) 给出 [0, 1]范围内的置信度。3) 提供简洁的理由（<100 字），依据背景档案↪和逻辑推理。避免猜测。

[背景档案]背景

[主张]主张

[输出要求]严格输出以下 JSON 对象：

"y": 0 或 1, // 0=正确，1=错误 "p": 数字, //浮点数，[0,1]，最多保留 2 位小数 "just": "..." // <100 个词，不要使用 Markdown

不要包含任何额外的键、说明文字或格式。↪

A.1.2 中心节点/根节点：单轮判断与提交的叶节点

用途： 中心节点读取主张并查看对等节点 JSON 报告；加权方式由公式（1）处理。

提示词。

你是中心（中心节点/根节点）事实核查代理。你将阅读主张↪，并查看一份对等节点报告列表。请给出你自己的最终判断。↪

[主张]主张

[对等节点报告]外围代理提供的 JSON 对象列表：PEER_JSON_LIST /* 每个元素的格式为：“智能体_id”：“Li_i”，“y”： 0|1，“p”： [0,1]，“just”：“ ”...↪ 请勿在输出中逐字复制这些内容。仅将其用作补充证据。↪*/

[任务]
1) 判断该主张正确（=0）还是错误（=1）。↪2) 给出一个位于 [0, 1]范围内的校准置信度 p。3) 给出简洁的理由（<100 词），并提及最↪重要的诊断性考量。避免盲从多数意见；应基于主张本身进行推理。↪

[输出要求]
严格按照以下 JSON 对象格式输出：

"y": 0 或 1, "p": 数字, "just": "..."

请勿添加任何额外的键、说明文字或格式。↪

A.2 分布式共识

A.2.1 初始判断

使用方式：每个智能体都在任何互动开始前做出初始决策。

提示词。

你是分布式共识环境中的自主智能体（没有中央↪控制器）。请给出你的初始判断。

[背景档案]背景

[主张]主张

[任务]
1) 输出 y，取值为 0 或 1，其中 0=正确， 1=错误。2) 输出一个校准置信度 p，取值范围为 [0,1]。3) 提供简短的理由（<100 词），并以事实和逻辑为依据。↪

[输出要求]
输出字符串 ictly the f以下 JSON 对象：

"agent_id": "AGENT_I D", "t": 0, "y": 0 或 1, "p": 数字 , "just": "..."

请勿包含任何额外的键、说明文字或格式。↪

A.2.2 交互轮次

用途：在每一轮 $t \geq 1$，智能体都会收到其先前状态和邻居报告。

提示词。

你正在参与一个多轮分布式共识过程。在第 t-1 轮，你将↪看到自己先前的立场以及邻居的报告。请提供你↪当前的立场，但请注意，最终状态由系统计算。↪

[上下文]
- 智能体：AGENT_ID
- 轮次：ROUND
- 你先前的状态 (t-1)：「y」：Y_PREV，「p」：P_PREV↪
- 邻居报告 (t-1)：NEIGHBOR_JSON_LIST

[任务]
1) 陈述你当前的立场（y 取值为 0 或 1，p 取值为 [0,1]）。↪2) 提供简洁的理由（<80 词），并提及决定性信号。↪3) 不要总结所有邻居；只提及那些会实质性改变你立场的内容。↪

[输出要求]
严格按照以下内容输出 JSON 对象：

"agent_id": "AGENT_I
D", "t": 轮次, "y": 0 或
1, "p": 数字, "just": "..."

请勿添加任何额外的键、文字或格式。↪

A.3 代理决策过程伪代码

以下伪代码描述了智能体的决策过程和信念更新机制：

```
# 初始化智能体的状态（信念，置信度）
智能体 = 初始化_智能体()

对于 t 在 range(T)：        # 循环对于 T 轮次
    的决策与信念更新
    对于 智能体 在 智能体集合中：
        # 步骤 1：评估主张使用 the 提示词（图 2）
        主张 = 获取_主张(t)
        背景_档案 =
            生成_背景_个人资料(主张)
        任务_描述 =
            生成_任务_描述()

        # 智能体生成一个二元判断并置信度分数
        y_i_t, p_i_t = 评估_主张(智能体, 背景_个人资料, 任务_描述)

        # 步骤 2：更新智能体的信念评分，使用更新规则（公式1）
        s_i_t_加_1 = 更新_信念(智能体, 智能体, y_i_t, p_i_)

        # 步骤 3：分配更新后的信念回到智能体
        智能体.信念_得分 =
            s_i_t_加_1

# 结束 of 模拟
```

列表1：代理决策过程的伪代码

B 模型设置

对于闭源大语言模型（GPT-3.5 和 GPT-4o），我们将所有运行的采样温度设为 0.7。我们使用 gpt-3.5-turbo-16k-0613 和 gpt-4o-1120。对于开源模型 llama3.3-70b-instruct，我们采用相同的温度以确保可比性；其他所有解码超参数均遵循

表4：大语言模型在 Snopes25 上的零样本性能。

模型	Acc.	精确率	Rec.	F1
GPT-3.5	0.64	0.62	0.60	0.61
GPT-4o	0.70	0.66	0.69	0.67
Llama3.3	0.67	0.63	0.66	0.65

Ollama。[1] 我们还在表4中报告了每个模型的零样本性能。

C 案例

我们展示 Snopes25数据集中的代表性案例研究，以说明系统的成功模式和失效模式。案例 C.1–C.4 展示了在不同拓扑结构下正确的集体判断，而案例 C.5 则突显了一种失效情形：分布式共识放大了最初不正确的解读。

C.1 分层集中式聚合（正确）

主张。 马桶未盖上马桶盖就冲水时，病原体可能会释放到空气中。

真实情况。正确。

拓扑结构与协议。一个由七个智能体组成的三级层级结构：一个根节点（R）、两个中间聚合器（M_L、M_R）和四个叶节点（L1–L4）。采用单轮协议：叶节点给出一次判断；中间聚合器汇总各自分配到的叶节点的判断；根节点作出最终决策。

叶节点输出（$t = 0$）。

"智能体_id":"L1","t":0,"y":0,"p":0.71,"just":"马桶羽流气溶胶化的证据 ↪表明，未盖上马桶盖时会发生空气传播。"↪"智能体_id":"L2","t":0,"y":0,"p":0.68,"just":"机 ↪理流体动力学和观察到的↪液滴形成现象表明，颗粒可能通过空气↪扩散。"↪"智能体_id":"L3","t":0,"y":0,"p":0.76,"just":"研 ↪究表明，冲水后颗粒计数会上升；马桶盖可以减少排放，但未盖上马桶盖会增加↪排放。"↪"智能体_id":"L4","t":0,"y":0,"p":0.73,"just":"报 ↪道的生物气溶胶与开盖冲水下的主张相符。"↪

中间聚合器。

[1] https://github.com/ollama/ollama

"y":0,"p":0.82,"just":"叶节点报告一致指出，↪不盖盖子冲水后会产生气溶胶颗粒。"↪"y":0,"p":0.80,"just":"多个叶节点提到颗粒计数和生物气溶胶增加；↪该主张得到支持。"↪

根节点决策。

"y":0,"p":0.86,"just":"两个子聚合器都对开盖冲水时的空气传播达成一致。↪"↪

C.2 星型集中聚合（正确）

主张。 冲马桶时若未盖上马桶盖，病原体可能释放到空气中。

真实情况。 正确。

拓扑结构与协议。 采用星型拓扑结构，由一个中心节点（H）和六个叶节点（L1–L6）组成。叶节点给出一次性判断；中心节点作出最终决策。

叶节点输出（$t = 0$）。

"agent_id":"L1","t":0,"y":0,"p":0.72,"just":"开盖冲水会产生含有↪气溶胶颗粒的马桶气溶胶羽流。"↪"agent_id":"L2","t":0,"y":0,"p":0.69,"just":"液滴和气溶胶形成支持无盖时的↪空气传播。"↪"agent_id":"L3","t":0,"y":0,"p":0.75,"just":"冲水后颗粒计数增加；马桶盖↪可减少排放。"↪"agent_id":"L4","t":0,"y":0,"p":0.71,"just":"生物气溶胶证据与开盖↪冲水的情况相符。"↪"agent_id":"L5","t":0,"y":0,"p":0.74,"just":"流体动力学表明，向上的羽流能够↪使微生物悬浮。"↪"agent_id":"L6","t":0,"y":0,"p":0.70,"just":"观测到的羽流高度支持空气↪扩散。"↪

中心节点决策。

"y":0,"p":0.86,"just":"所有叶节点都对羽流和气溶胶证据达成共识；该主张↪正确。"↪

C.3 环形分布式共识（2 个邻居，正确）

主张。 马桶在未盖上盖子的情况下冲水时，病原体可能会释放到空气中。

真实情况。正确。

拓扑结构和结果。 一个由七个智能体组成的环形结构，每个智能体都与两个邻居相连。系统在第$t = 3$轮次达成一致共识（$y = 0$）。

初始判断（$t = 0$）。

"agent_id":"A1","t":0,"y":0,"p":0.62,"just":"马桶羽流研究表明会释放气溶胶。"↪"agent_id":"A2","t":0,"y":1,"p":0.58,"just":"证据似乎不一。"↪"agent_id":"A3","t":0,"y":0,"p":0.65,"just":"冲水后颗粒数量有所增加。"↪"agent_id":"A4","t":0,"y":1,"p":0.55,"just":"效应大小可能有限。"↪"agent_id":"A5","t":0,"y":0,"p":0.60,"just":"流体动力学支持向上的羽流形成。"↪"agent_id":"A6","t":0,"y":0,"p":0.63,"just":"生物气溶胶报告与开盖冲水↪相符。"↪"agent_id":"A7","t":0,"y":1,"p":0.57,"just":"先前证据似乎尚无定论。"↪

在 $t = 1$ 和 $t = 2$轮次中，局部邻居互动逐渐使最初持怀疑态度的智能体转向多数立场。

共识（$t = 3$）。

"agent_id":"A1","t":3,"y":0,"p":0.74,"just":"邻域完全认同关于羽流的证据。"↪"agent_id":"A2","t":3,"y":0,"p":0.73,"just":"持续的认同足以证明正确。"↪"agent_id":"A3","t":3,"y":0,"p":0.75,"just":"证据在各轮次中始终一致。"↪"agent_id":"A4","t":3,"y":0,"p":0.70,"just":"邻居数据消除了先前的不确定性。"↪"agent_id":"A5","t":3,"y":0,"p":0.74,"just":"趋同的气溶胶观测。"↪"agent_id":"A6","t":3,"y":0,"p":0.75,"just":"支持性研究维持了正确结论。"↪"agent_id":"A7","t":3,"y":0,"p":0.73,"just":"累积的局部证据证实了该主张。"↪

C.4 完全图分布式共识（6 个邻居，失败案例）

主张。 当蜘蛛察觉到危险时，它们会跑向人群寻求保护。

真实情况。 错误。

拓扑结构与结果。 一个拥有七个智能体（每个有六个邻居）的完全图。尽管初始存在意见分歧，该系统在第 $t = 3$轮次达到了一致共识（$y = 0$，正确），从而得出了自信但错误的判断。

初始判断（$t = 0$）。

"agent_id":"A1","t":0,"y":1,"p":0.63,"just":"蜘蛛通常会避开人类；这种做法似乎↪不太可能。"↪

"智能体_id":"A2","t":0,"y":0,"p":0.58,"just":"因此，一些轶事性叙述表明，避难处位于↪大型物体附近。"↪"智能体_id":"A3","t":0,"y":1,"p":0.61,"just":"典型的反应是远离振动↪源。"↪"智能体_id":"A4","t":0,"y":0,"p":0.55,"just":"向静止的大质量物体移动可能会减少↪暴露。"↪"智能体_id":"A5","t":0,"y":1,"p":0.60,"just":"民族志文献强调回避↪行为。"↪"智能体_id":"A6","t":0,"y":0,"p":0.57,"just":"寻求避难处可能会无意中与↪人类所在位置一致。"↪"智能体_id":"A7","t":0,"y":1,"p":0.59,"just":"现有线索表明应撤退，而非接近。"↪

在第 $t = 1$ 和第 $t = 2$ 轮中，反复接触这种看似合理但依据薄弱的“最近庇护处”解释，会提升人们对其可信度的感知，尽管它得到的实证支持有限。

错误共识（$t = 3$）。

"agent_id":"A1","t":3,"y":0,"p":0.80,"just":"最近庇护处的解释说明了表面上的↪接近行为。"↪"agent_id":"A2","t":3,"y":0,"p":0.72,"just":"趋同的解释支持正确。"↪"agent_id":"A3","t":3,"y":0,"p":0.71,"just":"一致的寻找庇护所行为验证了这种解释。"↪"agent_id":"A4","t":3,"y":0,"p":0.74,"just":"气流和振动梯度提供了可信的避难处。"↪"agent_id":"A5","t":3,"y":0,"p":0.79,"just":"最近庇护处启发式解释了表面上的寻求保护行为。"↪"agent_id":"A6","t":3,"y":0,"p":0.75,"just":"网络中的一致意见使这一结论合理。"↪"agent_id":"A7","t":3,"y":0,"p":0.73,"just":"累积的报告支持为寻求保护而接近的行为。"↪

此案例说明，在高度连接网络中，强烈的从众压力如何放大一个连贯却错误的叙事，进而导致自信地错误分类。

D 指标的形式化定义

D.1 集中式聚合指标

令 $y^\star \in \{0, 1\}$ 表示给定主张的真实标签，令 y_c 表示中心节点（中心节点或根节点）给出的最终判断。

中心准确率（CA）。中心准确率衡量中心节点所作集体决策的正确性：

$$\mathrm{CA} = \mathbb{I}[y_c = y^\star], \tag{3}$$

其中， $\mathbb{I}[\cdot]$ 为指示函数。CA 直接反映枢纽介导聚合的可靠性，并将中心能力的影响与外围多样性的影响区分开来。

外围准确率 (PA)。令 $\mathcal{P}$ 表示外围（非中心）智能体集合。外围准确率定义为

$$\mathrm{PA} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbb{I}[y_i = y^\star], \tag{4}$$

用于衡量独立于聚合结果的个体层面平均正确率。

中心—外围一致性 (CPC)。为量化对中心决策的即时从众，我们定义

$$\mathrm{CPC} = \frac{1}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \mathbb{I}[y_i = y_c]. \tag{5}$$

CPC 衡量外围代理与枢纽的判断保持一致的程度，而不考虑判断是否正确，从而将从众强度与准确率区分开来。

D.2 分布式共识指标

在分布式共识中，智能体会在轮次 $t = 0$、.....、$T_{\max}$ 中迭代更新其判断。令$y_i^{(t)}$ 表示智能体 i在轮次 t 的判断。

最终准确率 (FA)。令 $\hat{}y$ 表示终止时的群体决策，定义为：若达成共识，则为一致标签；否则为 $T_{\max}$ 时的多数标签。最终准确率定义为

$$\mathrm{FA} = \mathbb{I}[\hat{y} = y^\star], \tag{6}$$

用于评估最终形成的集体结果是否与真实情况相符。

达成共识时间(达成共识时间)。达成共识时间定义为

$$\mathrm{TTC} = \min \left\{ t : y_1^{(t)} = y_2^{(t)} = \cdots = y_N^{(t)} \right\}, \tag{7}$$

若未能在$T_{\max}$内达成一致同意，则该指标未定义。TTC 反映了网络连通性和从众强度所带来的收敛效率。

一致性指数（CI）。在每一轮 t，从众指数定义为

$$\mathrm{CI}^{(t)} = \max\left(\begin{array}{c}\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[y_i^{(t)}=1\right],\\ \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left[y_i^{(t)}=0\right]\end{array}\right) \tag{8}$$

对于二元判断，$\mathrm{CI}^{(t)} \in [0.5, 1]$，其中 $\mathrm{CI}^{(t)} = 1$ 表示一致同意，$\mathrm{CI}^{(t)} = 0.5$对应最大意见分歧。

平均一致性指数（ACI）。为总结随时间变化的从众动态，我们报告平均 CI：

$$\mathrm{ACI} = \frac{1}{T_{\max}}\sum_{t=1}^{T_{\max}}\mathrm{CI}^{(t)}, \tag{9}$$

该指标反映了交互轮次中整体趋于一致的倾向，而非仅反映最终状态。