递归集体智能：多智能体大语言模型系统中等预算阶梯上的涌现

Lixuan He[1] Shengqu Cai[1] Fang Wu[1] Shikang Zheng[2] LiweiJiang[3] Manling Li[4]Yi Ma[2] Li Fei-Fei[1] Leonidas Guibas[1] Gordon Wetzstein[1] Yejin Choi[1] Yining Hong[1]

[1]斯坦福大学 [2]加州大学伯克利分校 [3]英伟达 [4]西北大学 lixuan27.github.io/RCI
github.com/lixuan27/RCI-official lixuan27.github.io/RCI/blog
lixuan27.github.io/RCI/playground

arXiv:2611.32606v1 [cs.MA] 24 Nov 2026

大语言模型越来越多地以智能体团队的形式开展工作，而人们期待团队能够解决任何单个成员都无法解决的问题。大多数评估会因团队优于单个智能体而给予团队认可，却没有考虑额外的计算开销：团队消耗的令牌可能是单个智能体的许多倍，而对独立答案进行简单投票，已经可以解释大部分已报告的增益。因此，仍有两个问题悬而未决：交互本身何时能让群体变得更聪明？群体能否通过学习自身工作来改进下一代？我们通过**等预算阶梯**回答这两个问题：该阶梯在三个层级上将群体与独立智能体进行比较，分别是优于单个智能体、优于计算量匹配的多数投票，以及优于计算量匹配的 pass@n。基于这一阶梯，我们将**递归集体智能（RCI）**定义为：群体无需教师或外部标签，仅凭自身输出，就能以相同成本让下一代变得更聪明。由此得出三项发现。第一，当交互促使信息在智能体之间流动时，集体智能会提升：转发拆分事实使群体在三种模型规模下均超越 pass@n，在 27B 规模下增幅约为三分之一。第二，组织方式会自发涌现：相同的 27B 智能体会形成枢纽，使准确率提高 18 个百分点；在提示词中编码其组织方式后，在一项描述性比较中，它们以不到一半的令牌预算达到了第3级。第三，群体会递归地自我改进：两个无标签代逐步攀升阶梯，第二个 9B 代超越了 pass@n，也超过了一个使用基础模型回合训练的对照组；在 PlanBench、HiddenBench 和 Hanabi 上，每一代在未见问题上的表现都优于前一代，这种优势持续了四到六代。因此，当群体转发有用信息的效率高于其在协调上消耗令牌的效率，并能从自身工作中学会这样做时，它就能胜过其成员。

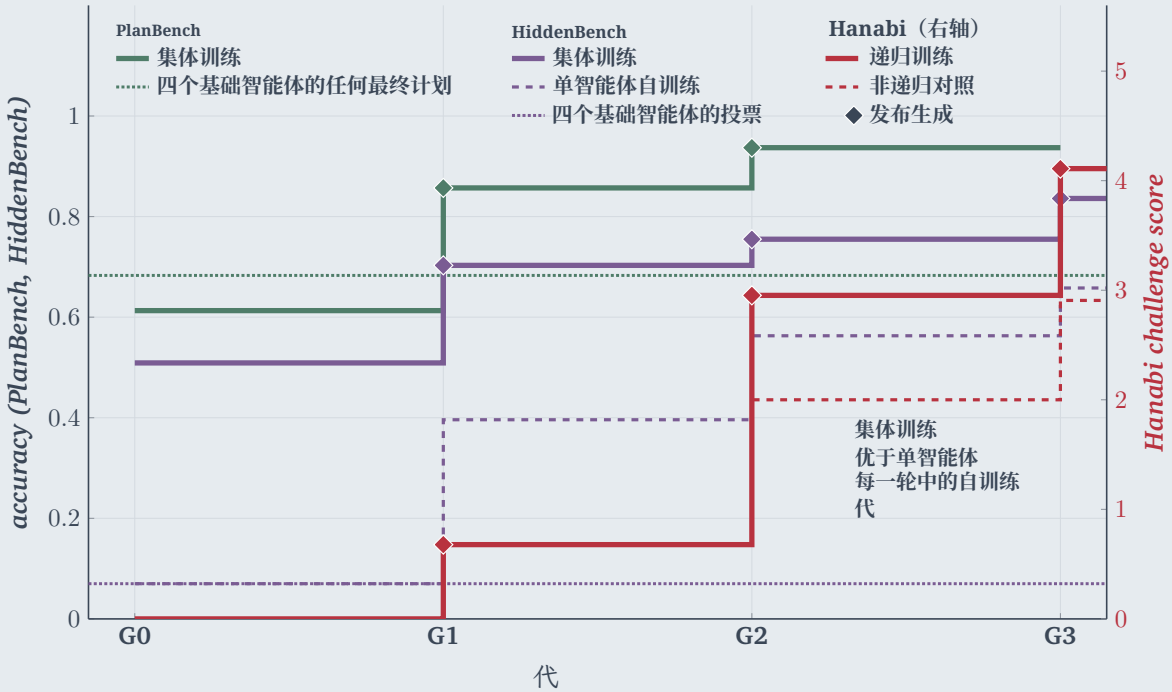


图 1：递归集体智能。每一代都基于前代经过验证的输出进行无标签训练；菱形标记发布节点，PlanBench 展示至 G2。从 G1 起，群体在 PlanBench 上的表现就超过了四个基础智能体的任何最终计划。在 HiddenBench 上，其表现远高于四个基础智能体的投票，也高于具有相同上限的单智能体自训练。在 Hanabi（右轴）上，递归训练的表现高于非递归对照（第7.4节）。

目录

1 引言 32 **相关工作 3**3 **从多智能体系统到集体智能 4**4 **集体智能的等预算阶梯 5**4.1 计算量匹配的集成与三个层级 .54.2 阶梯上的组织方式 .64.3 自我改进与 Shapley 贡献 .64.4 跨代门控递归 .7**5 群体何时比各自独立工作的成员更聪明？ 8**5.1 实验设置 .85.2 在共享信息的情况下，先前的多智能体方法并未优于多数投票 .85.3 转发拆分事实达到第3级 .8**6 涌现式组织及其协调成本 9**6.1 相同的智能体构建的枢纽与中继节点 .96.2 准确率、成本与协调开销 .11

7 群体能否自我改进？ 13

7.1 标准编写的提示词（测试时）.137.2 在经验证的交互轮次上训练.137.3 根据群体的结果训练.137.4 在 PlanBench、HiddenBench 和 Hanabi 上进行递归.137.5 等预算阶梯上的 HiddenBench 各代.16

8 讨论与结论 16

A 附录概述 22B 证明 22C 等预算阶梯上的评分 22D 隐藏档案协议与稀缺注意力 23E 协调令牌核算 24F 递归实验 24G 递归命名规则 26H 共享黑板与智能体测试框架 26I SplitBench、矛盾审计与 DocSweep 27J 评估协议与基线 28K 补充结果 29K.1 阶梯的支持性结果 .29K.2 Shapley 贡献详述 .29**L 在 PlanBench、HiddenBench 和 Hanabi 上进行递归训练 31M 计算资源与可复现性 32**

电子邮件： lixuanhe@stanford.edu

1. 引言

集体智能是指群体解决问题的能力优于其成员各自单独解决问题的能力，它可以作为群体本身的一种属性来衡量 [24, 66]。如今，语言模型以群体形式协作，可能具备这种属性。同一模型的多个副本通过共享记忆相互辩论或基于彼此的结果继续推进，并给出一个联合答案 [14, 21, 40, 44, 62]。这些系统带来两项前景：群体能够解决超出个体智能体能力范围的问题，并且能够通过从自身工作中学习而递归改进。

然而，要检验这些承诺，需要进行恰当的比较。在独立样本上进行多数投票时，每个样本都会产生自身的计算成本，这本身就已经优于单个智能体 [31, 64], , 但其中许多改进被归因于多智能体辩论 [11]。因此，超越单个智能体并不一定证明了交互的益处，而可能仅仅反映了更高的计算成本。当信息分布在不同智能体之间时，交互的价值尤为重要。然而，讨论往往聚焦于所有成员共享的信息，却忽视了只有一名成员知晓的事实 [56], , 语言模型群体中也观察到了同样的模式 [33]。与此同时，自我改进方法依赖外部反馈，并单独评估模型 [54, 70]。因此，仍有两个问题尚未解决：在相同预算下，交互本身何时能使群体比其成员更聪明；以及群体能否在没有教师的情况下，让其下一代变得更聪明?

我们用同一尺度回答这两个问题。人类集体智能研究会将群体表现与独立工作的个体进行比较，其中最强的基线是：只要群体中有任一个体答对，就视为群体答对，这与大语言模型评估中的 pass@n 类似 [30, 38, 57]。我们将这一框架扩展到语言模型智能体，并进一步考虑计算成本。等预算阶梯在相同令牌预算下，将群体表现与三个基线进行比较：单个智能体、多数投票，以及 pass@n （图2）。由于 pass@n为独立答案之间的任何选择方式设定了上限，达到第 3 级便表明，交互带来的增益超越了答案选择。进一步将边际分解为准确率和成本，可以揭示这些增益究竟源于答案质量提高，还是令牌效率提升。基于这一阶梯，我们进一步定义了**递归集体智能（RCI）**：群体自身的输出在没有教师或外部标签的情况下，能够提升其下一代。

基于 这一框架，我们开展了大量实验，取得三项主要贡献和一项发现 s:

• **等预算阶梯**。我们提出集体智能的三个层级：表现优于单个智能体、优于计算资源匹配的多数投票，以及优于 pass@n。我们证明，pass@n为独立答案选择设定了上限，并发现现有方法在共享信息任务上无法胜过计算资源匹配的投票，而 forwarding splitfacts 在 27B 规模上超出 pass@n约三分之一。

• **涌现式组织与协调**。 相同的智能体会自发形成枢纽和中继节点，使准确率提升18个百分点。在描述性对比中，将其组织方式编码到提示词中，可让群体以不到一半的令牌预算达到第3级。

• **递归集体智能**。 在我们的隐藏画像基准测试 SplitBench 以及另外三个基准测试上，群体通过对自身经过验证的输出进行无标签训练来实现自我提升。在 SplitBench 上，两个代逐步攀升阶梯：第二个 9B 代超过了 pass@n ，也超过了基于基础模型轮次训练的对照组；这两项改进还转移到了官方 HiddenBench 任务上。在 PlanBench、HiddenBench 和 Hanabi 上，每一代在未见问题上的表现都超过前代，且这一趋势持续四至六代；当信息分散在各智能体之间时，集体训练的表现优于单智能体自训练。

2. 相关工作

评估语言模型智能体群体。 人们通常认为，多智能体系统的表现优于单个智能体。我们认为，这种提升的部分原因仅仅是增加了计算量。对独立样本进行多数投票，本身就能提升单个智能体的表现 [8, 31, 64]。抽取的样本越多，其中至少有一个正确的可能性就越大 [6]。相比之下，多智能体辩论很少能胜过多数投票，或使用优质提示词的单个智能体 [11, 55, 63, 73]。模块化人工智能系统也引出了同样的问题：多个模型共同工作是否优于每个模型单独工作 [16]。因此，公平的评估应在相同计算量下，将群体与独立智能体进行比较。当信息分散在各个智能体之间时，这一点尤为重要，因为此时群体只有在智能体共享各自的私有事实后才能成功，而语言模型群体往往会跳过这一步 [33]。我们的阶梯明确呈现了这种比较。

组织语言模型智能体。 多智能体系统以多种方式组织智能体：智能体进行辩论 [14], 汇总分层提案 [62], 或通过有向图进行协调 [44], 、借助中央协调器 [12, 48] 或共享黑板 [21, 22, 40]。其他研究则在答案达成一致后停止抽样 [2, 32] ，或分析多智能体系统失败的原因 [7]。相关基准测试通过模型之间交换的内容 [15] 以及长周期软件项目来衡量协作 [53]。这些系统预先固定了组织方式，因此我们想知道，完全相同的智能体能否自行形成组织方式。

训练智能体与集体。 智能体强化学习通过组相对策略梯度，训练单一模型以实现可验证的结果 [28, 51, 68, 72], 而蒸馏则从教师那里学习 [1, 20] 并从成功运行中进行模仿 [9, 13, 41]。多智能体扩展利用验证器奖励辩论者，或共享团队奖励 [23, 35, 36, 43, 76]。由于共享奖励对每个智能体都相同，它掩盖了每个智能体所作的贡献。协作强化学习通过差分奖励、价值分解和夏普利值来分配这一信用 [18, 46, 52, 59, 61, 65]；我们的 Shapley 贡献将同一理念应用于共享单一策略的智能体。

递归自我改进。 哥德尔机等自指系统会改进用于改进自身的过程 [49]。语言模型通过在自身的推理依据、经过筛选的答案、自分配奖励或多数投票结果上进行训练来改进 [25, 54, 69, 70, 77]。智能体支架会演化其提示词 [17], 其改进器 [71] 或自身的代码 [74], , 模型群体则在自身的辩论上进行微调 [58]或在成功的团队轨迹上进行微调 [75]。这些方法使用标签或验证器为答案评分，并分别评估每个模型。相比之下，我们的代从由群体自身协议验证的回合中学习；只有当某项改进在相同成本下优于独立智能体时，阶梯才会将其贡献归于该群体。

3. 从多智能体系统到集体智能

第 2 节回顾的系统，主要区别在于信息如何在智能体之间流动。本文始终使用以下术语。 群体 是指一组运行相同模型、使用相同工具并返回一个联合答案的智能体。其 组织方式 是决定每个智能体写入和读取哪些消息的规则，例如对独立答案进行多数投票、辩论、由中央协调器协调，或使用共享黑板。其 令牌预算 是其在一项任务上花费的完成令牌数量，以相同令牌所能支付的 n_B 独立智能体数量表示。无论采用何种组织方式，增加智能体都会提高令牌预算；我们认为，有三项约束决定了这笔更大的预算能带来什么。

信息结构。 当每个智能体都能看到整个任务时，增加智能体只会从相同分布中增加样本，其价值受限于从独立答案中进行最佳选择的效果，即 pass@n (命题 1)。当决定性事实分散在不同智能体之间时，解决任务就需要将这些事实结合起来，而这只有通过交互才能实现 [33, 56]。因此，我们分别报告这两种情形。

相关错误。 同一模型的副本会产生相关错误。对它们的答案进行多数投票会很快达到饱和 [6,8]。此外，交换消息往往会使智能体与多数意见保持一致，即使少数意见才是正确答案 [5, 11, 37]。

协调成本。 智能体写入或读取的每条消息都会消耗令牌，而独立智能体本可以将这些令牌用于各自的尝试；此外，一个智能体的失败还可能通过其他智能体的消息传播开来 [7]。因此，只有当交互传递的信息价值超过其成本时，交互才是值得的。

集体智能。 借鉴对人类群体的研究，我们将集体智能与同一批个体分开工作时的表现进行比较 [30, 38, 57, 66], , 并固定令牌预算。一个群体在 具备集体智能 时，意味着它在相同令牌预算下超越独立智能体；其最强形式还超过它们的 pass@n，而任何从独立答案中进行的选择都无法达到这一点。若相同的智能体能在局部规则下自行形成这种组织方式，例如限制每个智能体可以读取的消息数量，则其组织方式是 涌现的 。当一个群体无需教师或外部标签，仅凭自身输出就能以相同成本使其下一代变得更聪明时，它便展现出 递归集体智能 。

这些主张也可能源于群体评估的方式。一个表现优于单个智能体的群体，可能只是消耗了更多令牌。同样，得分高于前代的某一代，可能只是在重复使用的评估所产生的噪声上被选中的。4 因此将每个主张转化为群体可能未能通过的一项测试。等预算阶梯在同一样本池上，将每个群体与计算量匹配的独立智能体进行比较，并将每个边际拆分为准确率和成本；只有当新一代在此前任何轮次都未使用过的任务上优于前代时，门控递归才会发布新一代。第5、6 和7 节将这些测试应用于交互、涌现和递归。

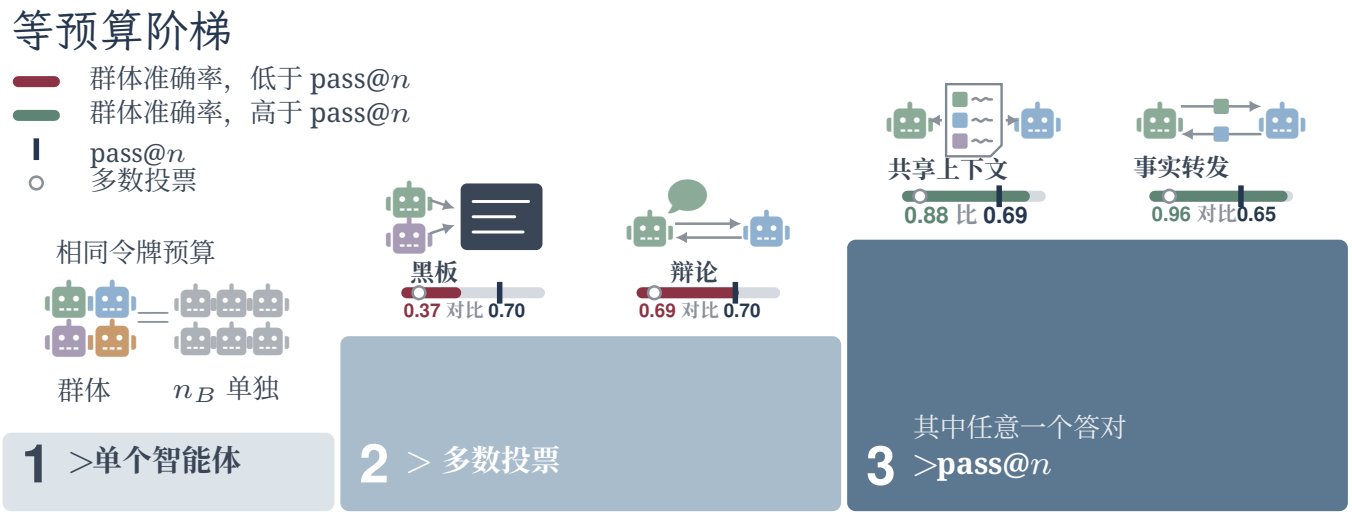


图2：等预算阶梯。我们通过三个层级定义等预算阶梯：对于令牌预算相当于运行 n_B 个独立智能体的群体，当其表现优于单个智能体时达到第1级，优于 n_B 个独立智能体的多数投票时达到第2级，超过它们的 pass@n_B 时达到第3级。HiddenBench 上的四种 Qwen3.8-27B 多智能体协议展示了这一阶梯，每种协议均归入其达到的最高层级。

4. 集体智能的等预算阶梯

4.1. 计算量匹配的集成与三个层级

一个群体由 N 个运行相同模型的成员组成。成员通过协议进行交互，群体最终提交一个答案，其准确率为 Y。作为参照，我们也将同一模型作为独立智能体运行。每个独立智能体接收一名成员所获得的信息，我们逐个席位抽取这些智能体组成集成。如果一次群体运行消耗了 k 个完成令牌，其预算为 $n_B = k/c_1$ 个智能体，其中 c_1 是一个独立智能体的平均成本。该阶梯只需要每个智能体的最终答案和令牌数，因此适用于任何模态的智能体，包括读取页面图像的智能体（第5.2节）。

对于由 n 个 V独立智能体组成的集成，$n(T)$ 表示它们多数投票的准确率，$n(n)$ 表示它们的 pass@ 准确率，即至少有一个智能体答对的概率。我们定义了集体智能的三个层级。

- **第1级**，$Y > V(1)$：群体的表现优于单个智能体，因此额外使用的令牌带来了更好的答案。
- **第2级**，$Y > V(n_B)$：群体的表现优于由 n_B 个独立智能体进行多数投票的结果。
- **第3级**，$Y > T(n_B)$：群体的表现超过 pass@n_B，因此它能给出任何独立答案组合都无法选出的正确答案。

按顺序检验各层级。若以问题为聚类单位的单侧配对检验在 0.025 的显著性水平上确认从第 1 级到 k 的每个层级，则群体达到第 k 级。表现低于单个智能体的群体处于第 0 级。我们称群体在第 2 级优于多数投票，在第 3 级优于 *pass@n*；预算明确时，我们将 pass@n 记作 pass@n_B。参照协议的三个属性解释了这些层级（证明见附录 B）。

命题 *1*（引用的结构）。设每个独立智能体以p_x的概率正确回答问题 x，且彼此独立。*(i)* $T(n) = 1 - \mathbb{E}_x[(1-p_x)^n]$ 是单调不减且凹的：每增加一个智能体，增量为$\mathbb{E}_x[p_x(1-p_x)^n]$。*(ii)* 任何承诺从 n 个独立智能体的答案中选择一个的规则，其正确率至多为 $T(n)$，即使它还使用这些智能体的文本、置信度或评审者的意见。因此，$T(n) \geq max\{V(1)\ V(n)\}$，而第*3*级意味着第*1*级和第*2*级。*(iii)* $T(n) \leq min\{1, n\,V(1)\}$，因此增加 $n-m$ 个智能体，*pass@*n 最多提高 $(n-m)\ V(1)$。

第（ii）部分将阶梯变成交互测试。从多数投票到学习型评审者，每一种选择规则都会从独立答案中选出一个，因此其准确率始终不高于相同成本下的 pass@n。因此，若群体的表现超过 pass@n，就必定是成员通过交换而得出了正确答案。第（iii）部分说明了何时可能得到这类答案。对于分散事实，单个智能体很少能答对，因此 pass@n 至多随预算线性增长。此时，汇集事实的群体可以胜过所有选择规则，因此其增益主要来自准确率。对于共享信息，$V(1)$

较大且通过 @n 会在少数几个智能体内趋于饱和，因此群体必须胜过一个几乎完美的选择器。

准确率与成本。考虑协议 X 和参照协议 R，它们针对同一样本池进行评测。边际的变化可分解为准确率的变化，以及由不同预算n导致的 pass@

$$\begin{aligned}&\big[Y_X - T(n_X)\big] - \big[Y_R - T(n_R)\big] \\ &\quad = \underbrace{\big(Y_X - Y_R\big)}_{\text{accuracy}} + \underbrace{\big[T(n_R) - T(n_X)\big]}_{\text{cost}}.\end{aligned} \tag{1}$$

命题 1 将这一核算恒等式转化为关于成本的表述。成本项是 $n_R - n_X$额外独立智能体所能带来的准确率增益。根据命题 1（iii），它至多为 （$n_R - n_X$） V（1），因此，只有当额外智能体仍能提高准确率时，更廉价的协议才会获得较大增益。X 通过准确率达到第3级，当其反事实边际 $Y_X - T$（n_R）（即在 R 的预算下的边际）被确认大于零时，便可实现这一点。否则，X 既需要更高的准确率，也需要更低的成本。

4.2. 阶梯上的组织

我们评估阶梯上的几种常见群体组织方式，图 5a 根据成员交换的内容将其归为七个类别（附录表8）。在辩论中，成员以自由文本形式交换答案和理由，进行两或三轮 [14]。在黑板中，成员在公共黑板上发布和读取消息 [21, 40]。证据协议则交换事实。在事实转发中，成员只转发自己的私有事实。共享上下文将每位成员的事实附加到共享记录中。仅事实黑板只接受转发的事实，而门控黑板要求成员读取每一条事实后才能发布分析。为研究自组织，我们为每位成员提供一条消息的阅读预算，该预算可以转移给另一名成员。收集到足够预算的成员可以读取每一条事实，并成为枢纽。对每位成员的预算设置上限，可以防止出现单一枢纽，因此群体必须构建中继节点。可选的规划轮次允许成员在读取之前就角色达成一致。设计型组织由人工分配这些角色。设计型中心将全部预算交给一名成员，而设计树则分配枢纽和中继角色。在共享语料库上，设计的聚合器让三名读取者分担文档，第四名成员汇总他们的发现。采用自编写提示词时，群体遵循由模型自行编写的组织指令。最后，令牌核算将每次模型调用标记为等待、协商或工作（附录 E）。

4.3. 自我改进与 Shapley 贡献

本节定义递归集体智能。群体可以通过两种方式自我改进：在测试时写入提示词，或训练其权重。我们还介绍了基于群体结果进行训练的信用分配规则。

递归。 当一个群体无需教师或外部标签，便能通过自身输出在同一阶梯上改进其下一代时，我们称其具备递归集体智能（图 3）。在测试时，模型会为其自身群体写入提示词。在权重训练中，每一代都根据其父代群体协议验证过的回合进行训练。这些主要是首轮发言，即成员在读取彼此内容之前写下的发言，并且必须涵盖成员的私有事实。第二代每个席位采样四个回合，并保留通过验证的最短回合。在 27B 中，首轮发言在标准辩论上下文中进行训练；4B 和 9B标准辩论运行还会在承诺轮次上进行训练（附录 F）。群体会进行两种辩论中的一种。 标准辩论 允许成员在“事实已被分割”的提示下自由讨论三轮，而三步辩论 将每个成员限制为三个步骤。在这两种辩论中，协议对每个回合的检查取代了答案标签。当以下三个条件同时满足时，我们称其为递归：第一代（G1）优于基础模型（G0），第二代（G2）优于 G1，并且 G2 击败了一个对照组，该对照组基于 G0 的回合进行训练。

Shapley 贡献。 当协议不检查个体回合时，采用同一策略的成员必须根据群体结果进行训练，而信用分配规则决定了该策略学到什么。在集合值任务中，群体交付其成员主张的并集$P_S = \bigcup_{i \in S} P_i$。联盟 S 在任务评分 $v(S) = s(P_S)$ 下的价值为 s，其中 $v(\emptyset) = 0$，而群体的奖励为 $R = v(\{1, \ldots, N\})$。假设群体持有 r 个不同声明，其中 h 个为真，在 t 个真实发现中。一名成员随后以校准置信度 q添加一项主张，即该主张为真的概率。

命题 2 (报告阈值)。 在召回率 h/t 下，该声明会提高每个 $q > 0$ 的期望得分。在 *F1* 分数 $2h/(r + t)$ 下，当且仅当 $q > \mathrm{F1}/2$ 时，该声明会提高得分。对于真实声明加 $1/t$ 分、虚假声明减 c/t 分的评分方式，当且仅当 $q > c/(1 + c)$ 时，该声明会提高得分。

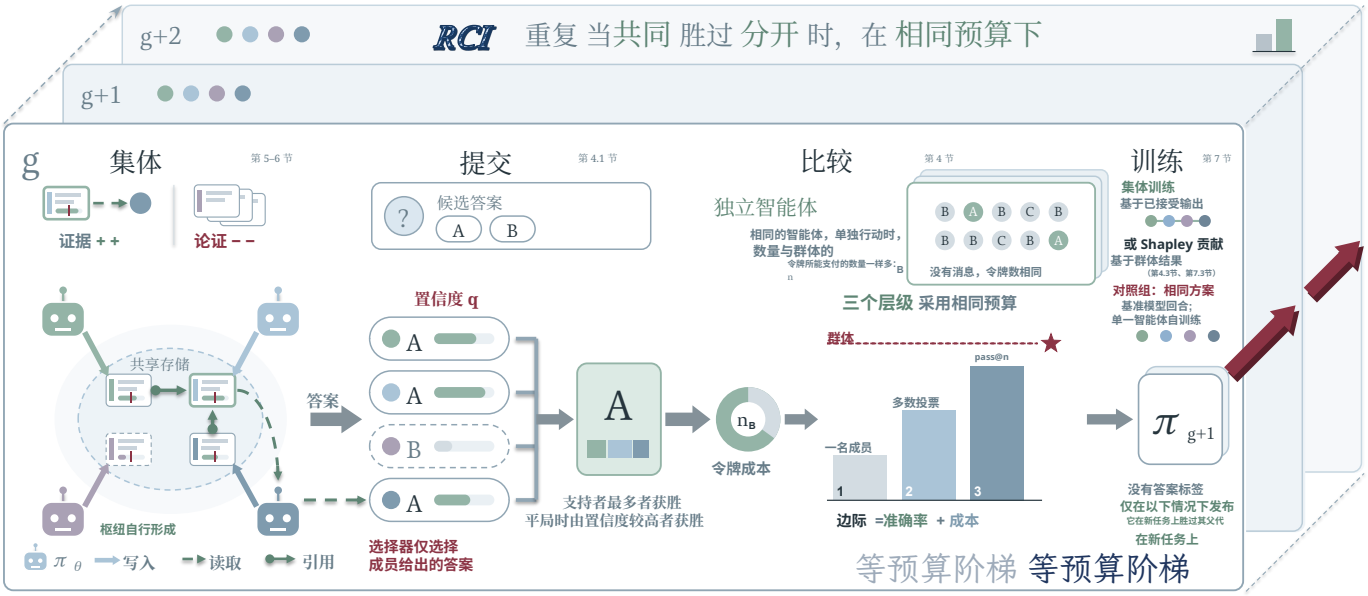


图3：一代的概览 g。集体：策略的副本 π_θ 向共享存储写入消息、读取并引用彼此的消息，还可以将自己的阅读预算交给另一位成员，因此枢纽会自发形成；证据有所帮助（++），而论据会误导（−−）。提交：每位成员都提交一个答案及其置信度 q，群体提交支持者最多的答案；若票数相同，则选择置信度较高的答案，因此选择器只会从成员给出的答案中进行选择。比较：记录中的完成令牌决定其成本 n_B，即这些令牌所能支付的独立智能体数量；相同的智能体在相同信息条件下单独运行，并将群体表现与 n_B 这些智能体的三个层级进行比较：单个成员、多数投票以及 pass@n，边际由准确率和成本两部分构成（公式1）。训练：下一代 π_{g+1} 基于群体自身协议接受的输出进行训练，不需要教师或答案标签（集体训练）；如果没有协议检查单轮结果，则基于群体的结果进行训练，并按每位成员获得的 Shapley 贡献支付（第4.3节）。通过对照组检验增益：一个对照组使用相同方案，基于基础模型的回合进行训练；在 HiddenBench 上，还会与单智能体自训练进行比较，即每个智能体都在自己的输出上训练。只有当候选项在新任务上的表现优于其父代时，才会发布（图4），随后循环重复。各面板标题标明了详细阐述每个步骤的章节。

Shapley 贡献 会向每位成员分配其在群体得分中的沙普利份额 [52]，

$$\phi_i = \sum_{S \subseteq \{1,\dots,N\} \setminus \{i\}} \frac{|S|!\,(N - |S| - 1)!}{N!} \big(v(S \cup \{i\}) - v(S)\big), \tag{2}$$

因此，多个成员报告的发现会在他们之间进行拆分，而错误的主张则计入其作者名下。每份正向共享的比例$\gamma = 0.5$ 随后会沿着引用传递给其消息所依据的成员，从而得到f_i。成员 i 获得信用 $r_i = \lambda R + (1 - \lambda) N f_i$，其中 $\lambda = 0.25$，训练采用裁剪群体相对策略梯度 [50, 68]（附录K.2）。

4.4. 跨代门控递归

两个无标签代表明，一个群体能够自我改进，这引出了改进是否会持续这一问题。在更长的序列中，我们看到两个风险。基于含噪评估进行选择，可能会发布一个仅仅匹配其前代的代，而每一代所学到的内容也可能少于前一代。因此，我们在 PlanBench、HiddenBench 和 Hanabi 上，将递归作为由多代组成的门控序列运行（图 4）。在这三个基准测试中，当每个发布生成在未见任务上都优于其前代时，我们称一次运行是递归的；而上述三个条件则定义了 SplitBench 上的递归。

一轮。当前一代使用自己的集体协议，在一组全新的训练任务上进行测试；自动检查会在没有人工编写的答案标签的情况下，接受或拒绝每项输出：包括 PlanBench 中的计划模拟器、HiddenBench 中与至少四名基础模型成员中有三名对每项事实给出的答案的一致程度，以及 Hanabi 中的游戏得分。候选项从当前一代的权重继续训练，主要使用其最新的已接受输出进行训练，并在选择集上排名。通过健康门槛的最佳候选项随后会与当前一代在以下任务上进行比较

确认任务，这些任务此前轮次均未使用，并按任务配对。只有当其配对增益 d 超过阈值（例如，在 PlanBench 中超过 $d \geq 1.645$ 个标准误）时，它才会成为下一代；否则该轮次为平坦的。每轮使用新的确认任务，可避免候选项因重复使用同一选择集所产生的噪声而被选中；以基础模型为基准设定的健康门槛则会淘汰令牌预算、答案格式或对照组任务准确率下降的候选项。

单智能体自训练。 为检验增益是否归功于群体，在 HiddenBench 上，一项运行（T3）增加了匹配实验：在相同的任务数和训练令牌上限下，分别使用每个智能体自己的单独输出训练该智能体；Hanabi 则增加了一项非递归对照实验，使用固定代的对局进行训练。这样一来，集体训练与单智能体自训练的唯一区别，就是训练数据是否来自群体交互。附录 L 列出了每项运行、对应的任务集及其发布规则。

5. 群体何时比各自分开的成员更聪明?

5.1. 实验设置

模型。 我们评估了四种通义千问模型（Qwen3.5-4B、Qwen3.5-9B、Qwen3.6-35B-A3B 和 Qwen3.8-27B [45, 67]）以及另外两个模型家族的模型（GLM-4.5-Air [19]和 gpt-oss-120b [42]）。所有模型均通过 vLLM [29], 提供服务，且每个群体使用一种模型。
评估模式。 三种模式的区别在于每个成员能看到什么。在信息共享模式下，每个成员都能看到完整的问题；我们使用 GPQA-Diamond [47], LongBench v2 [3], MMLongBench-Doc [39],BrowseComp-Plus [10], SWE-bench Verified [27]和 FATE-M [26]。在信息拆分 模式下，决定性事实分散在各个成员之间。HiddenBench 包含 65 个此类决策任务 [33]。表 2 中的阶梯实验和自由预算组织实验使用全部 65 项任务；提示词和预算上限实验及其作为比较对象的自由预算参照实验，则使用恰好包含四条私有事实的 56 项任务（附录 D）。我们还以相同格式构建了 SplitBench。它保留了由 Qwen3.8-27B 编写的 1,500 项任务中的 697 项；成员只有汇集事实才能解答这些任务，其中 501 项用于训练，196 项留作测试（附录 I）。在共享语料库 模式下，矛盾审计和 DocSweep 将答案分散在多份文档中，每个成员都可以打开这些文档（附录 I）。

5.2. 共享信息条件下，既有多智能体方法无法超越多数投票

图 5c 比较了三种既有方法与计算量匹配的多数投票在六个共享信息基准测试上的表现。使用通义千问3.5-9B智能体时，我们重新实现了多智能体辩论 [14], DeLM，其中智能体共享一个黑板， [40],以及 AOrchestra，其中协调器分派子任务 [48]。表1进一步评估了由四个通义千问3.8-27B智能体组成的群体在 GPQA-Diamond 和 DocSweep 上的表现。

三种方法在每个基准测试上的表现都不高于计算量匹配的投票，因此仍低于第2级。由于每个智能体都已经掌握证据，一条消息大多只是在重复其读者已知的信息。对于页面图像，这一结论同样成立：在 MMLongBench-Doc 上，辩论的表现与投票相当，而 DeLM 和 AOrchestra 的表现低于投票（附录表12）。在 GPQA-Diamond 上，27B黑板和辩论仍处于第0级（表1）。正如图6a 所示，充满信心的智能体会在投票中胜过唯一正确的智能体。

DocSweep 展示了哪些因素会改变这一结果。答案分散在一个共享语料库中，没有任何智能体能够完整读取该语料库。自行组织的智能体会打开相互重叠的文档，因此表现优于投票，但未能达到 pass@n。自编写提示词或精心设计的聚合器则会将文档分配给阅读者，这些群体的表现超过了 pass@n (表1)。因此，当每个智能体都能看到完整的问题时，先前的方法仍停留在投票水平；只有从成员的答案中进行选择的审阅者才能更进一步（附录表12）。在共享语料库中，分工阅读有所助益，有组织的群体表现超过了 pass@n。

5.3. 转发拆分事实达到第3级

表2 和图5b 展示了 HiddenBench 上的结果。四名 Qwen3.8-27B 成员各自掌握部分决定性事实，而共享事实则指向一个诱饵。因此，单独工作的智能体会跟随诱饵，导致 pass@n 保持较低水平，而汇集事实的群体则能超越这一水平。出于同样的原因，第2级的门槛很低：各自只掌握部分事实的智能体大多会投票选择诱饵。每种证据协议都达到了第3级。辩论协议的表现大多优于投票，并停留在第2级。它们的消息传递基于部分事实得出结论，读者会采纳这些结论，而不是自己缺少的事实。第三轮中加入事实被拆分的提示后，辩论在更高预算下提升到了第3级。事实转发在 9B、27B 和 35B-A3B 下均达到第3级。在 27B 下，四名各自掌握部分事实的智能体甚至比等预算全信息智能体更准确（0.959 对 0.942，表2）。事实转发的表现也优于共享上下文，

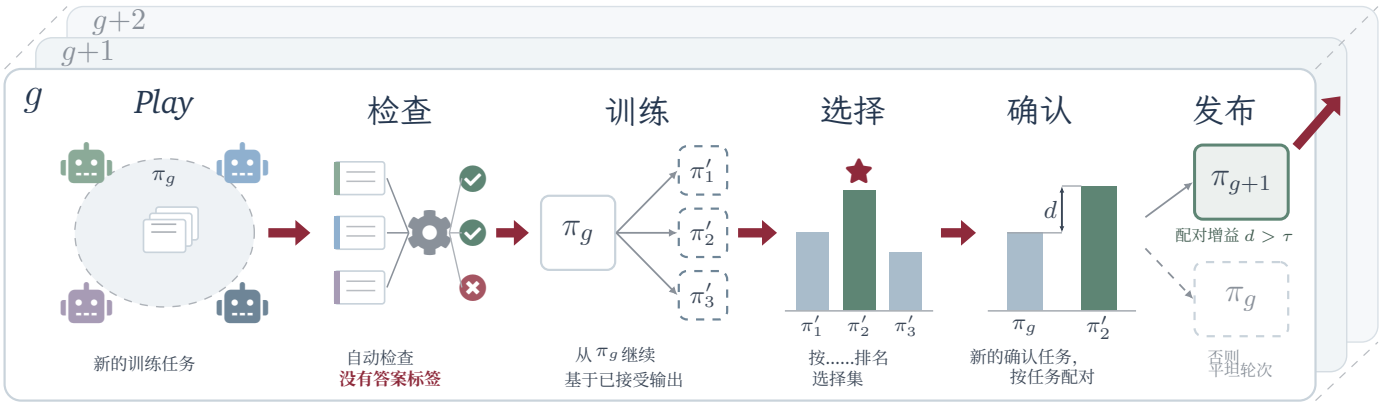


图4：跨代门控递归。每一轮中，第 g 代使用其集体协议执行新的训练任务，自动检查（计划模拟器、协议检查或游戏得分）会在不使用人工编写的答案标签的情况下，接受或拒绝其输出。候选项从 π_g 代的已接受输出继续进行，并在选择集上排名；排名最佳的候选项会与 π_g 代在新的确认任务上按任务配对进行比较，仅当其配对增益 d 超过阈值 τ 时，才会作为第 $g{+}1$ 代发布。否则，本轮结果为平坦的。

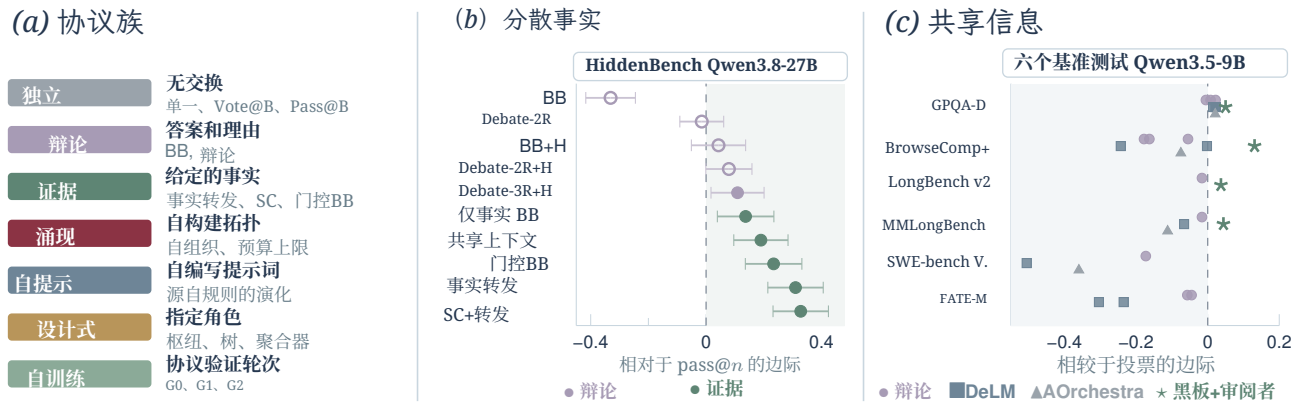


图 5：协议族及其在阶梯上的位置。(a) 我们根据成员交换的内容将协议分为七个家族，每个家族始终使用同一种颜色。黑板（BB）是所有成员共享的留言板。共享上下文（SC）会将每个成员的事实附加到记录中。Vote@B 和 Pass@B 分别是群体令牌预算下独立智能体的多数投票和通过率@n 。 **(b)** 对于分散事实，证据 协议的表现优于通过率@n。实心标记达到第3级，条形表示95%区间， +H表示事实经过分发的提示。**(c)** 在共享信息的情况下，无论是在文本上还是在 MMLongBench-Doc 的页面图像上，辩论、DeLM 和 AOrchestra 的表现都从未超过投票。星号表示我们的黑板配置，其中27B审阅器会从9B成员的答案中进行选择；该配置在运行过的四个基准测试上都优于投票。

评论附加的位置　　d 传递的事实中包含论据，即使是掌握所有事实的智能体也会被误导（图6b　　）.

6. 涌现式组织及其协调成本

6.1. 由相同智能体构建的枢纽与中继节点

当智能体自行构建某种结构时，我们称其为涌现结构。在表 3中，HiddenBench 上的四个相同的 Qwen3.8-27B 智能体可能各自读取一条消息，并且每个智能体都可以将这一阅读预算转移给另一个智能体。面板（a）评估三种设置：自由预算、自编写提示词，以及防止任何智能体读取所有事实的上限。图 7 展示了个体记录。

在自由预算下，大多数智能体会将预算转出。在约三分之一的记录中，单个智能体会收集到足够的预算，从而读取所有事实并成为枢纽。该枢纽的回答如同获得每项事实的单个智能体，因此准确率比不进行转移的群体高出18个百分点。规划轮次强化了这一行为，使群体接近设计型中心。在预算上限下，智能体转而构建中继节点，主要是两层树（图 7b）；这些中继节点的表现优于只能转发自身事实的群体。所有这些结构的表现都优于投票，并且仍处于第2级。

表1：共享信息和共享语料库。四个 Qwen3.8-27B 智能体。 n_B 统计群体的令牌为多少个智能体支付了费用，边际为准确率减去 pass@n_B。按顺序测试的各层级均以经确认的边际优势击败单个智能体、多数投票和 pass@n ； 第 0 层未能击败单个智能体。

协议	Acc.	n_B	边际	Lvl.
GPQA-Diamond：所有智能体使用相同的问题				
黑板	0.909	4.22	−0.039	0
辩论 ，2 轮次	0.899	4.24	−0.049	0
DocSweep：答案分散在共享语料库中				
自组织	0.200	4.93	+0.026	2
自编写提示词	0.861	4.02	+0.706	3
设计的聚合器	**0.930**	4.56	**+0.759**	3

表2：分散事实的结果。HiddenBench 上的四个 Qwen3.8-27B 智能体分别持有部分决定性事实；在带有提示的行中，智能体会被告知事实是分散的。传递事实可达到第3级，且事实转发的表现超过等预算全信息智能体。

协议	Acc.	n_B	边际	Lvl.
辩论：成员交换答案和理由				
黑板	0.369	8.6	−0.331	2
辩论 ，2 轮次	0.688	8.8	−0.015	2
辩论 ，3 轮次 +提示	0.841	11.1	+0.109	3
证据：成员传递各自的事实				
共享上下文	0.881	7.9	+0.190	3
事实转发	0.959	6.2	+0.310	3
共享上下文 + 转发	**0.972**	6.0	**+0.328**	3
全信息智能体，相同预算	*0.942*			

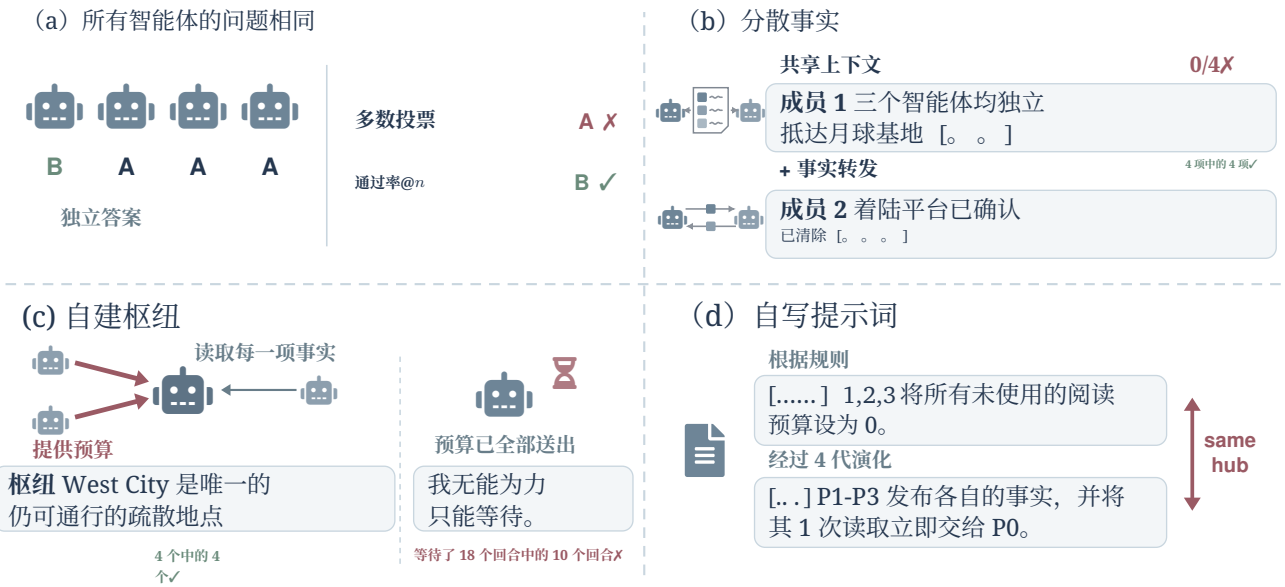


图6：定性结果，通义千问 Qwen3.8-27B 的记录。**(a)** 多数投票推翻了那位答对的成员的答案。**(b)**在共享上下文中，一项论证误导了那位掌握所有事实的成员；而转发的事实则让四位成员都得出了答案。**(c)** 成员将阅读预算交给一个枢纽，而另一群体的一名成员随后等待自己的枢纽。**(d)** 仅根据规则撰写的提示词就已经描述了同一个枢纽。方框中的文本为逐字引用。

它们的预算解释了为何其层级不会再提高。一个自组织群体会花费约21个独立智能体的令牌，预算是设计型中心的两倍，因为其智能体会等待枢纽并协商转移（附录表9b）。

表3：涌现式组织。 HiddenBench 上的相同 Qwen3.8-27B 智能体，每个智能体读取一条消息。**（a）**成员可以转让自己的预算（自由预算，全部 65 项任务），或者为每个智能体设置预算上限，使其无法形成单一枢纽（预算上限，包含四条私有事实的 56 项任务）。自由预算运行的提示词行使用来自另一批次的样本池在这 56 项任务上运行，因此其层级仅具有描述性；树的层级在多重检验校正前成立。**（b）**记录包含枢纽，或包含转移但没有规划轮次；同时展示在 65 项任务上允许转移后准确率的变化。

组织方式		自由预算 Acc.	n_B	边际	Lvl.	预算上限 Acc.	n_B	边际	Lvl.
成员	无传输或中继	0.496	20.3	−0.300	2	0.571	37.0	−0.267	2
	自组织	0.677	21.3	−0.124	2	0.689	40.7	−0.159	2
	+ 规划轮次	0.822	18.3	+0.037	2	0.677	44.4	−0.180	2
提示词	根据规则	**0.879**	9.2	**+0.205**	3	0.732	40.2	−0.115	2
	演化	0.862	9.7	+0.180	3	**0.840**	25.0	+0.043	2
设计	枢纽／树	0.863	11.3	+0.130	3	0.802	10.8	**+0.107**	3

（b）模型	Hubs	转移	Gain
Qwen3.5-4B	0.000	0.024	+0.027
Qwen3.5-9B	0.000	0.008	−0.001
Qwen3.6-35B-A3B	0.000	0.004	+0.025
Qwen3.8-27B	**0.345**	**0.611**	**+0.181**
GLM-4.5-Air	0.000	0.017	+0.025
gpt-oss-120b	0.045	0.372	+0.036

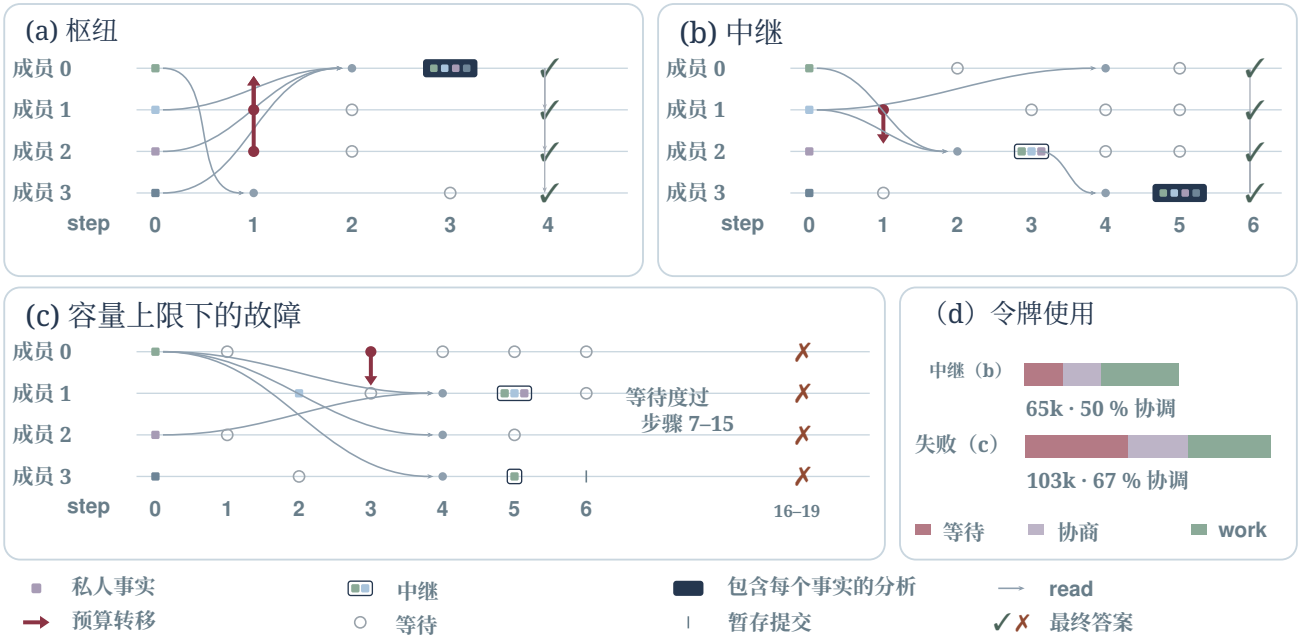


图7：逐条读取并追踪自行构建的结构。每个面板展示 HiddenBench 上一条通义千问Qwen3.8-27B记录的演变过程。面板 (a) 追踪图6c的枢纽，(b) 展示预算上限下的中继，(c) 展示容量上限下的故障，而 (d) 比较 (b) 和 (c) 的令牌使用量。

6.2. 准确率、成本与协调开销

图9 将公式（1）应用于每个协议族，并将各协议与同一样本池上的参照协议进行比较（附录表9）。纵轴为反事实边际 $Y_X - T(n_R)$，即参考预算下的边际。**证据**协议即使在其参考协议的预算下，仍能确认高于 pass@n，因此其增益仅由准确率解释。**设计式**和**自提示**组织方式也提升了准确率，但只有在自身较低的预算下才超过 pass@n，因此它们既需要更高的准确率，也需要更低的成本。**涌现**结构的准确率高于其参照协议，但在两种预算下，其超过 pass@n 的边际都未达到显著水平。自组织群体将约三分之二的令牌用于协调（附录表9b）。在自由预算群体中，大部分令牌用于协商转移，其余则用于等待枢纽，

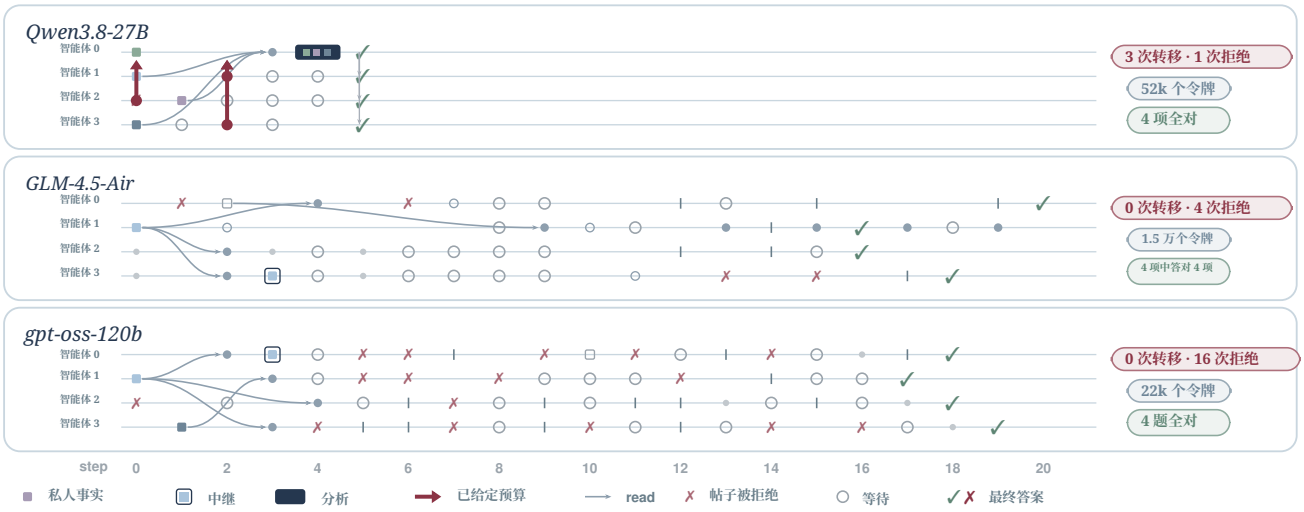


图8：一项任务，三类模型。OneHiddenBench 任务采用带有规划轮次的自组织方式（泳道：智能体；筹码：私有事实）。通义千问 Qwen3.8-27B 智能体将阅读预算交给单个智能体，由其在六步内发布分析；而 GLM-4.5-Air 和 gpt-oss-120b 智能体则转发单一事实。

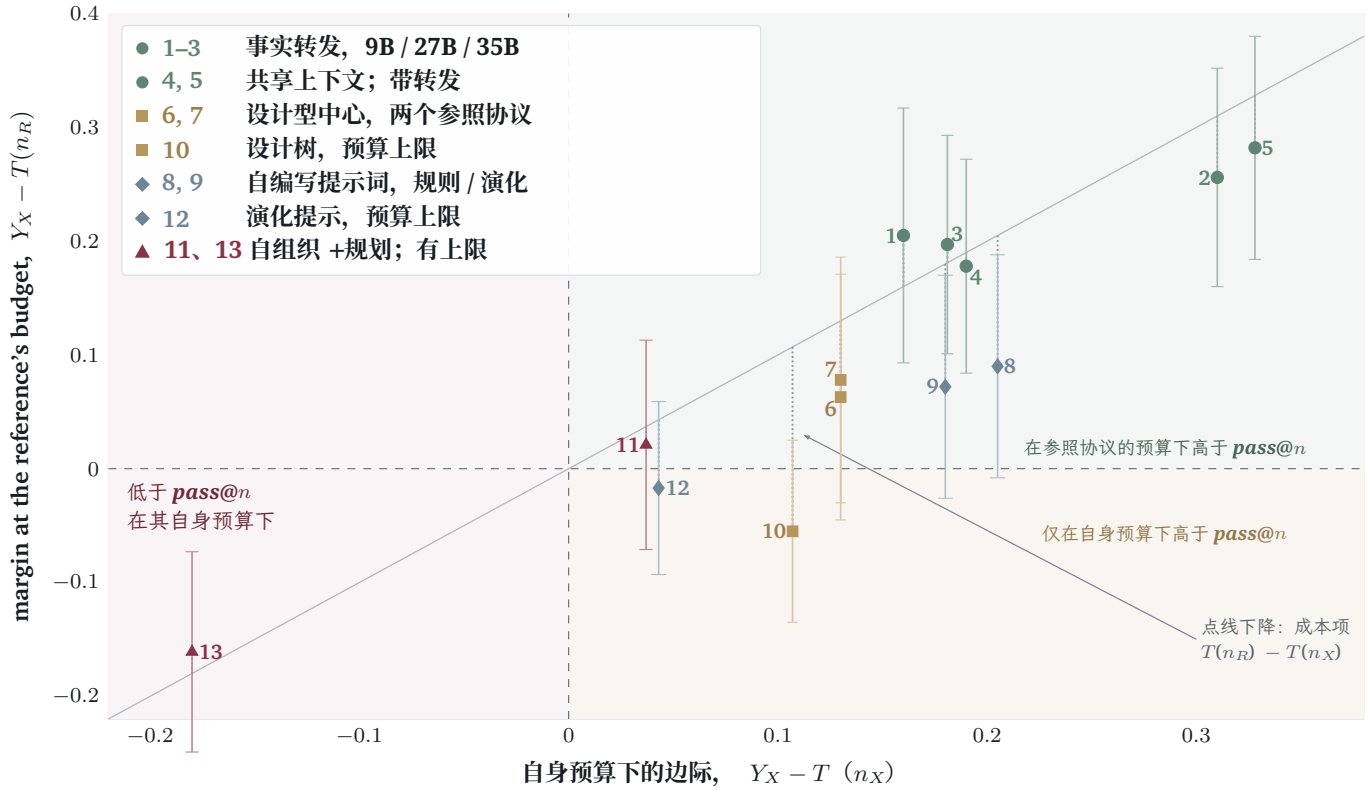


图9：准确率与达到第3级的成本。每个点表示同一样本池中协议 X 与其参照协议 R 的对比（附录表9；误差线：95%）。对角线表示相同预算，从对角线到各点的点线下降表示方程1中的成本项，即 pass@n 在两个预算之间的变化；当 X 的花费低于 R 时，该点位于对角线下方。区域根据点估计绘制。证据协议即使在参考预算下也确认高于 pass@n；设计式和自提示组织方式 6 至 10 仅在其自身较低预算下确认高于该值；受限提示 12 和涌现结构在任一预算下均未确认高于该值。

其余令牌则用于等待枢纽。若没有这段等待，枢纽将超过 pass@n。

跨模型家族。在通义千问模型中，只有27B会形成枢纽；GLM-4.5-Air和gpt-oss-120b智能体也很少形成枢纽，而且允许转移并未显著改变任一模型家族的准确率（表3b）；图8 展示了每个模型家族中一个包含规划轮次的记录。

7. 群体能自我改进吗?

7.1. 测试时的自写提示词

在测试时，模型会为自己的群体编写提示词。它可以根据规则编写提示词，也可以在无标签记录上经过四代演化提示词。该提示词将27B群体从第2级提升到第3级（描述性结果，表3a，图6d）。即使是根据规则编写的提示词，描述的枢纽也与演化提示词相同；遵循该提示词的群体每条记录消耗的令牌不到自组织群体的一半。当规则没有明确规定时，经验会有所帮助。在预算上限下，演化提示词使仅依照规则编写的提示词的边际提高了0.16。拥有相同权重的监督者也可以根据无标签记录重写27B团队的提示词，从而使结果提高0.11（附录 F）。

7.2. 在经验证的交互轮次上训练

表4 报告无标签代。每一代自训练代都在父代群体针对 501 项 SplitBench 任务的协议验证轮次上训练，并在 196 项留出任务上评分。

第一代在所有规模上都提升了群体表现。在 9B 和 4B 规模下，边际分别提高 0.225 和 0.421；在 27B 规模下，边际则从 -0.009 提高到 $+0.201$。第二代在三个规模上都进一步提升。在 9B 规模下，第二代保留了通过所需的最短交互轮次，将首次调用的步数减半。随后，群体的表现超过 pass@n ，并在第一个训练种子上以 0.192 的优势超过对照组。汇总另外四个种子后，相较于对照组的增益为 0.115，且每一对比较中均为正值；准确率是增益的主要来源（附录 F）。因此，递归的三个条件在 9B 规模下均成立。在 27B 规模下，第二步的提升来自成本：预算从 9.6 个智能体降至 6.4 个，同时准确率保持在父代水平。两步在标准辩论和官方 HiddenBench 任务上也均成立（附录 F）。

增益归于群体（附录表10）。在标准辩论中，第一步使经过训练的 9B 智能体单独作答的准确率与之前持平，而在两代成员单独都能解决的任务上，其群体获得了 0.177 的增益。在三步辩论中，第二代的表现保持在其自身成员 pass@n 之上。

7.3. 训练群体的结果

当一项任务只根据群体的联合交付物进行评分时，信用分配规则决定共享策略学到什么。正如阶梯的预算一样，它必须让每个智能体为其新增的主张承担代价。表 5 在矛盾审计中检验了这一点；该审计根据召回率、精确率和 F1 值，对群体声称的矛盾对进行评分。奖励个体成功或团队成功，会促使智能体报告更多配对，并使群体的精确率崩溃。命题 2 解释了原因：智能体自身的评分会奖励低置信度主张，而群体要为这些主张造成的错误买单。Shapley 贡献则将每条错误主张的责任归于其作者。只有该方法能将留出集精确率提升到未训练组之上，并且它在每项指标上都表现最佳。

7.4. 递归　在 PlanBench、HiddenBench 和 H Hanabi

递归应当超越两代，并在不同任务中持续有效。因此，我们将第 4.4 节中的门控递归应用于三项测试不同能力的基准测试：PlanBench 中的计划生成 [60], 、HiddenBench 中的分散事实汇聚 [33] ，以及 Hanabi 中隐藏信息条件下的协作 [4]。每次运行都从四个通义千问 Qwen3.5-9B 智能体开始；每一代都使用前代的输出进行训练，这些输出须经计划模拟器接受、与基础模型的全信息答案一致，或获得相应游戏得分，且不需要人工编写的答案标签。在这三项基准测试中，每个发布的生成在未见过的确认任务上的表现都优于前代，连续四到六代皆如此（图 10）。

PlanBench。四轮均产生一代，群体在选择集上的准确率从 0.613 升至 0.950（表6）。智能体与群体共同进步：单个智能体的初稿和最终计划在每一代中都有所改善；从 G1 开始，只要四个智能体中有任何一个找到有效计划，群体几乎都会在每项任务上提交有效计划。随着准确率接近上限，经确认的增益逐渐缩小。由于后续各代中达到令牌上限的调用较少，准确率的提升部分可能源于思考时间缩短；因此，表格在列出每代准确率的同时，也报告了令牌数。

表4：无标签代。在196个SplitBench留出任务上，基础模型智能体相较于pass@n的边际。对照组按照G2的方式，在G0的回合上进行训练； Δ 表示代际变化。9B标准辩论的步骤在评估种子14至17上读取，其结果复现了随机种子10至13上的读数（附录 F）。

	相较于pass@n的边际			步骤 Δ	
模型	G0	G1	G2	G0→G1	G1→G2
三步辩论					
Qwen3.5-9B	−0.253	−0.028	**+0.159**	+0.225	+0.187
Qwen3.8-27B	−0.009	+0.201	**+0.284**	+0.210	+0.083
标准辩论					
Qwen3.5-9B	−0.228	−0.090	+0.098	+0.138	+0.188
Qwen3.5-4B	−0.498	−0.077	+0.117	+0.421	+0.194

对照组	边际	G2减去对照组
Qwen3.5-9B，3步	−0.033	**+0.192**
Qwen3.8-27B，3步	+0.016	+0.268

表 5：信用决定训练是否有助于群体。在共享语料矛盾审计任务上训练 40 步后， N Qwen3.5-4B 智能体的留出集召回率、F1 值和精确率（%）；依据群体的精确配对 F1 进行排名；基线：自身成功（GRPO）、模仿胜者（RAFT）、一名受训者和团队奖励（最佳项加下划线）。

	$N = 4$			$N = 8$		
方法	召回率	F1	精确率	召回率	F1	精确率
未训练	6.2	10.5	33.3	13.0	19.5	33.9
基线						
自身成功（GRPO）	5.5	7.7	12.2	9.5	10.2	9.4
模仿获胜（RAFT）	7.8	12.3	30.9	12.2	16.3	22.7
一名受训者	5.2	8.4	25.6	10.9	16.3	30.2
团队奖励	8.3	10.9	15.3	13.7	15.8	16.0
Shapley 贡献（我们的方法）	**9.2**	**15.9**	**58.4**	**14.2**	**20.1**	**49.7**

随着后续各代中达到令牌上限的调用减少，准确率的提升部分可能源于思考时间缩短；因此，表格在列出每代准确率的同时，也报告了令牌数。

表 6：PlanBench（Blocksworld 规划生成）。在由 300 个 Blocksworld 问题组成的选择集上运行四个 Qwen3.5-9B 智能体；每一代都基于前代进行训练，使用群体自身经计划模拟器接受的计划。群体提交其智能体匿名评审评分最高的最终计划。单个智能体评分一个智能体的首份草案和最终计划，而四者任一则在任一智能体的最终计划有效时计为解决该问题。截断表示达到 8,192 词元上限的调用所占比例。 d 表示在一个全新的 800 个问题确认切片上，相对于前代的配对增益及其标准误。

		单个智能体						
代	**群体**	**草案**	**最终**	**四者之一**	**截断**	**步数 / 每次调用的最优令牌数**		d **(标准误)**
G0（基准）	0.613	0.371	0.423	0.683	0.598	1.017	7,416	–
G1	0.857	0.480	0.610	0.860	0.447	1.018	6,802	+0.295 (0.020)
G2	0.937	0.591	0.707	0.937	0.343	1.019	6,651	+0.080 (0.014)
G3	0.940	0.630	0.752	0.947	0.297	1.018	6,523	+0.024 (0.012)
G4	**0.950**	**0.672**	**0.773**	**0.950**	0.265	1.022	6,541	+0.021 (0.010)

HiddenBench。 T3 和 T3k 两次运行在训练数据的构建方式以及候选项的步骤大小上有所不同（附录 L）。集体训练使群体在 T3k 的五代中从 0.509 提升至 0.886，在 T3 的三代中提升至 0.836（图 11a）。在相同上限下进行的单智能体自训练同样有所改进，而且在每一代中，集体的表现都远高于以这种方式训练的四个智能体。当事实分散在各个智能体之间时，该群体因此能够学会其智能体单独无法学会的内容。作为参照，图 11b 展示了我们使用同一基础模型在 65 项官方任务上评估的协议，图 11c 展示了基准测试的作者报告的群体。

Hanabi。 基准智能体在每局游戏中都会耗尽所有生命，得分为零。递归训练的六代在新牌局上的表现均优于前代，挑战得分升至 25 分中的 4.5 分（表 7）。对照实验改为在固定代的游戏上训练，而不是在自身产生的游戏上训练；经过两代后有所提升，随后稳定在 2.9 分左右，远低于

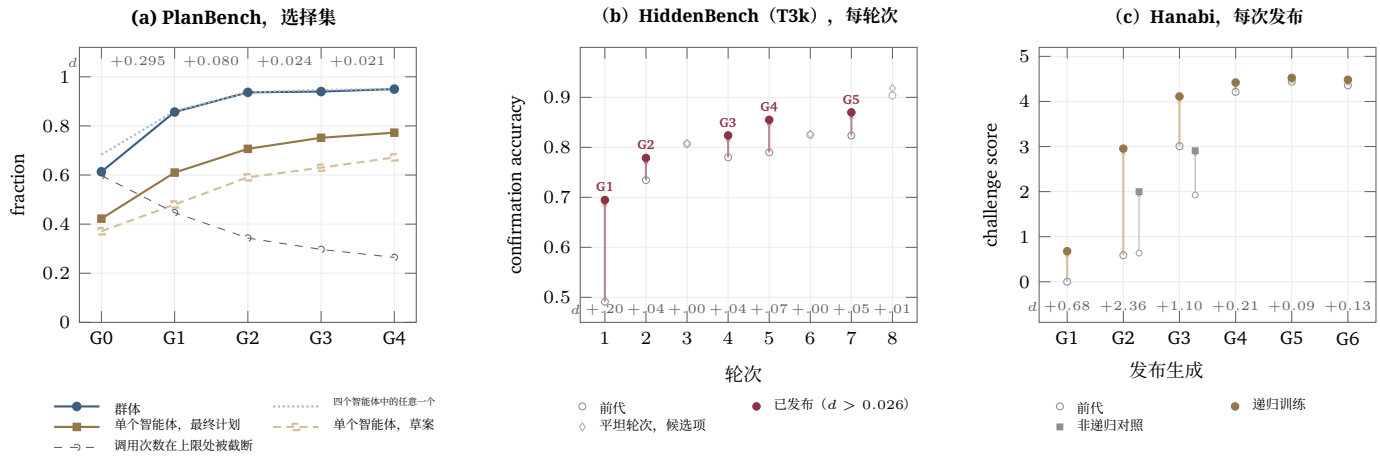


图10：PlanBench、HiddenBench 和 Hanabi 上的递归。（**a**）PlanBench，选择集上的第 0 至第 4 代：群体、预言机（四个智能体中任一正确即计为解决问题）、单个智能体的最终计划和初稿，以及达到令牌上限的调用比例。顶行给出每一代相对于其前代在全新 800 个问题的确认切片上的配对增益 d。（**b**）HiddenBench（T3k），全部八轮：每个候选项在该轮的全新确认种子上与当前代进行比较。五轮超过阈值并发布 G1 至 G5；在三轮平坦轮次中，候选项未超过阈值，因此保留当前代。（**c**）Hanabi：每个发布生成在 600 个新牌局上与其前代进行比较，以及非递归对照实验的两代。配对在不同的确认集上测量，且不同轮次之间不相连。

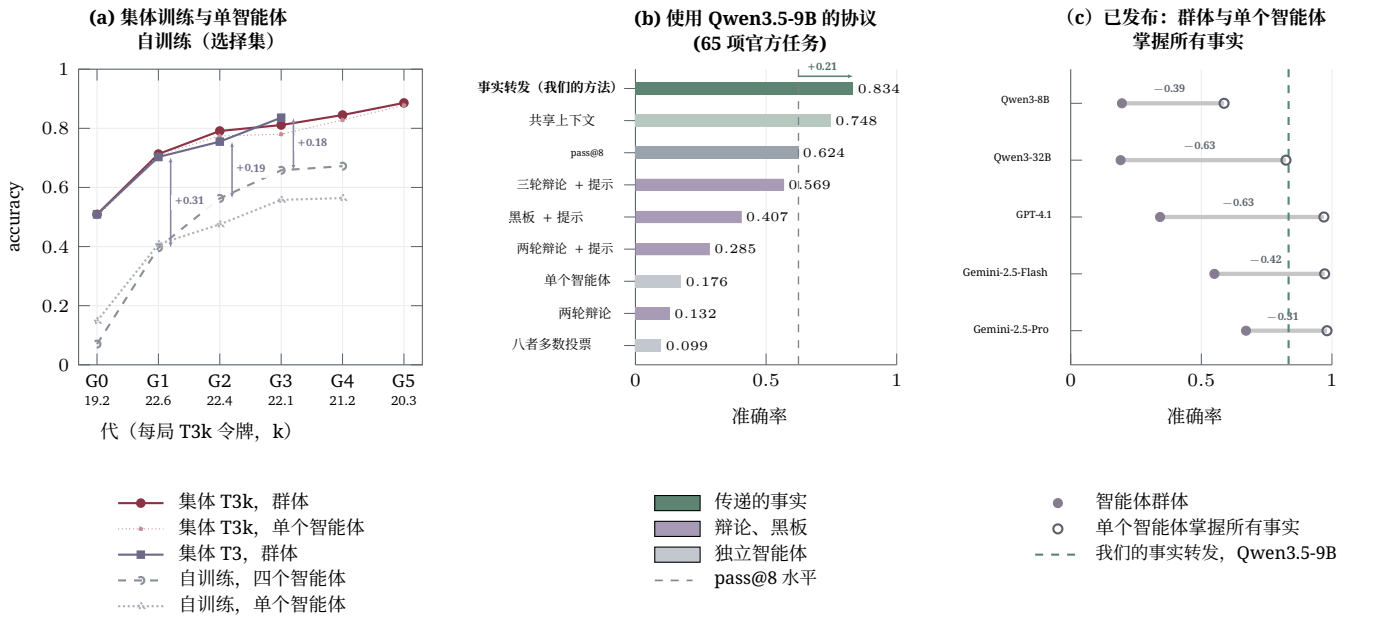


图 11：HiddenBench。（**a**）四个 Qwen3.5-9B 智能体在包含 200 个生成的隐藏档案任务的选择集上进行三轮辩论，并使用两个随机种子运行。运行 T3 和 T3k 中，每一代都基于群体经验证的交互轮次训练其前代；单个智能体表示辩论后该智能体自己的答案，代数下方的数字表示每场 T3k 对局的令牌数。单智能体自训练在相同上限下，使用每个智能体自己的单独回合进行训练，并分别按四个独立智能体和单个智能体进行评分；箭头表示 T3 与其匹配的自训练运行之间的差距。（**b**）第 5 节阶梯实验中，使用相同基础模型在 65 项官方任务上的协议；虚线表示八个独立智能体的 pass@8，事实转发的结果高出 0.21；附录表9列出了其等预算边际。（**c**）基准测试的作者报告的 65 项官方任务结果 [33]，摘自其图3：各群体与一个掌握所有事实的单个智能体进行比较，并显示群体损失的准确率。

递归训练。早期候选项最初只是通过少打纸牌来减少生命损失，因为训练过滤器移除了所有不确定的出牌。

表 7：Hanabi。四个 Qwen3.5-9B 智能体进行合作式纸牌游戏；挑战得分统计打出的纸牌数量，当团队失去全部生命时得分为零（最高 25 分）。每轮在 600 个新牌局上将候选项与其前代进行比较，因此每一行都将某一代与其前代在相同牌局上进行配对。递归训练让每一代都在其前代的精英对局上进行训练。对照实验从 G1 分支出来，并在固定代的对局上进行训练，而不是在自身对局上进行训练。

Run	代	挑战得分		增益（标准误）	已出的纸牌	失去所有生命
		代	前代			
递归训练						
	G0（基准）	0.000	–	–	1.362	1.000
	G1	0.678	0.000	+0.678 (0.045)	1.213	0.558
	G2	2.953	0.590	+2.363 (0.075)	2.953	0.000
	G3	4.110	3.008	+1.102 (0.087)	4.110	0.000
	G4	4.420	4.212	+0.208 (0.077)	4.420	0.000
	G5	**4.523**	4.437	+0.087 (0.074)	4.523	0.000
	G6	4.482	4.357	+0.125 (0.075)	4.482	0.000
非递归对照						
	G2	2.000	0.635	+1.365 (0.069)	2.003	0.002
	G3	2.907	1.925	+0.982 (0.085)	2.917	0.005

7.5. 等预算阶梯上的 HiddenBench 代

准确率单独来看，已经表明每一代都比前代有所进步。为了了解这种进步在成本相等时意味着什么，我们将 HiddenBench 的每一代都放到阶梯上进行比较（图 12）。对于每一代，使用该代自身模型的 24 个独立智能体回答全部 200 项选择任务；四个席位各使用六个随机种子。每个智能体都持有共享事实和一个席位的私人事实，就像群体中的成员一样。随后，以该代自身的预算对群体与这些智能体进行评估。

基准群体胜过单个智能体，也胜过其独立智能体的投票，因此从第2级起步，且比 pass@n_B 低 0.15。从第一代训练后的模型开始，每一代都超过 pass@n_B，并达到第3级。边际从 G1 的 $+0.06$ 增长到 G5 的 $+0.23$，与此同时，预算从相当于 13.9 个独立智能体降至 7.9 个。训练后智能体的 pass@n_B 在每一代都保持在 0.65 左右，而单个智能体独自工作时的成绩仍低于 0.18，因此训练后的智能体仍需要彼此掌握的事实。由此，递归训练通过交互沿阶梯向上攀升：每一代都学会汇集成员掌握的信息，且成本低于基准群体。

这套解读框架由四项特性构成。由于每个独立智能体只掌握四分之一的私有事实，此处的第3级衡量的是群体汇集分布式事实的能力。参照答案是基础模型的全信息共识，与训练时相同，因此，全信息智能体若以自己的答案作为评判对象，就无法充当参照。200项任务构成了训练期间使用的选择集，因此群体的准确率偏乐观，而独立智能体则是在全新的任务上运行。后续几代通常会单独作答，且答案格式无法解析；按此解读，这些答案会被计为错误。如果允许这些答案弃权，则G3至G5的投票结果会提高，而各级结果均不变。

8. 讨论与结论

上述实验将三个通常分开研究的问题联系起来：群体何时优于其中的智能体、群体如何组织自身，以及群体能否让下一代变得更好。由此得出四点经验，每一点都将关于多智能体系统的定性论述转化为经过测量的条件。

信息结构同时决定推理和训练。 测试时，交互只有在传递n 任何单个智能体都不掌握的事实时，才能超越pass@5（第5节）。训练中也再次出现了同一条件。在事实分散于各智能体之间的HiddenBench上，集体训练在每一代都远胜单智能体自训练，尽管两者使用相同的上限（第7.4节）。当集体的工作包含任何个体都无法独自收集的信息时，集体就能学到个体无法独自学到的内容；对于缺乏这种结构的任务，增加智能体只会增加样本，而其效果受到阶梯上限的约束。

递归受到前沿剩余提升空间的限制。 群体从其有时能够解决的问题中学习。随着群体不断改进，更多问题变得容易，留在其前沿的问题则越来越少。在 PlanBench 上，随着群体接近选择集的顶端，已确认的增益从第一代到第四代逐渐缩小；在 Hanabi 上，后续步骤相较于最初的步骤也很小。因此，自我改进的集体需要一个能够在其前沿不断提供新问题的来源，就像它需要训练规则一样

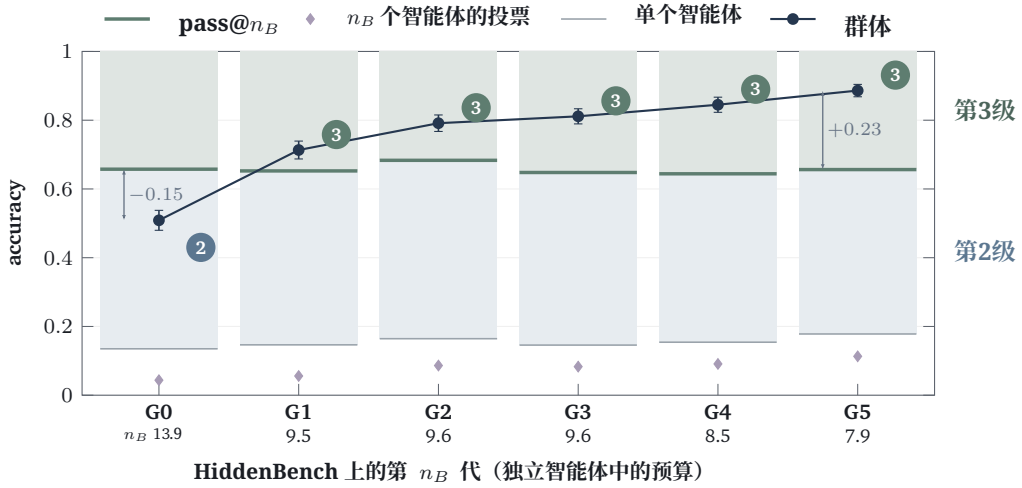


图 12：HiddenBench 各代模型沿等预算阶梯不断攀升。每一代模型都在 200项选择任务上接受评估，并与同一模型的 24 个独立智能体进行比较；这些智能体各自持有共享事实和一个席位的私有事实。阴影步骤标示各级别的范围：高于单个智能体和第2级的投票（蓝色），以及高于n_B 第3级的 pass@（绿色）。基准群体从第2级起步，而每一代训练后的模型都以更低的预算达到第3级。独立智能体组成了一个全新的样本池，因此单个智能体的准确率与图11a.略有不同。

同样需要训练规则。

小幅进步需要新的证据。从当前一代继续迭代，可在三个基准测试上持续发布新版本，而后续步骤的幅度较小。因此，每次发布决策都依据一部分此前轮次从未见过的数据。若在多轮中重复使用确认集，就会抬高当初据此选出的那一代模型的得分，真正的改进也就很少能通过门槛。

增益可能来自更省成本的行为。阶梯将每项边际拆分为准确率通道和成本通道（公式 1）。不同代之间也需要采用同样的拆分。在 Hanabi 中，候选项一开始通过少打纸牌来减少失去生命的次数；在 PlanBench 中，随着每次调用消耗的令牌减少，各代模型截断的调用也更少。这些变化都能提高得分，却不能证明群体的规划能力更强。将每一代的成本与准确率一同报告，并筛除那些能够获得奖励分数、却未体现相应技能的行为所对应的训练数据，就能将递归改进与更省成本的行为区分开来。

结论。 等预算阶梯揭示了群体何时比各自分开工作的成员更聪明：只有当交互为智能体提供了它们独自无法获取的信息时，交互才有价值。在共享信息的情况下，既有多智能体方法的表现仍停留在计算量匹配的投票水平，而转发拆分事实则在三个模型规模上超越了 pass@n 。因此，只有当群体交换的信息价值高于其用于协调所消耗的令牌时，群体的表现才会优于其成员。群体还可以从自身的回合中学习：两代无标签的 9B 生成依次取得提升，并胜过对照组，从而证实了递归集体智能。因此，多智能体系统应在相同成本下进行评估，并接受训练以高效传递信息。

跨越更多代的递归遵循同一原则。在 PlanBench、HiddenBench 和 Hanabi 上，每一代都连续四到六代优于其前代；当信息分散在多个智能体之间时，集体训练的表现优于单智能体自我训练。持续递归取决于能否持续获得群体前沿的未解决问题，以及每次发布时能否获得新的证据。因此，我们认为，自我改进的集体应依托新的问题来源。每次发布都应通过未见问题的检验，且每次比较都应确保成本相同。

引用

[1] Rishabh Agarwal、Nino Vieillard、Yongchao Zhou、Piotr Stanczyk、Sabela Ramos、Matthieu Geist 和 Olivier Bachem。《语言模型的在策略蒸馏：从自生成错误中学习》。发表于国际学习表征会议（ICLR），2024 年。[2] Pranjal Aggarwal、Aman Madaan、Yiming Yang 和 Mausam。《逐步采样：用于高效推理和编码的自适应一致性》。发表于自然语言处理实证方法会议（EMNLP），第 12375–12396 页，2023 年。[3] Yushi Bai、Shangqing Tu、Jiajie Zhang、Hao Peng、Xiaozhi Wang、Xin Lv、Shulin Cao、Jiazheng Xu、Lei Hou、Yuxiao Dong、Jie Tang 和 Juanzi Li。《LongBench v2：迈向对真实长上下文多任务的深入理解与推理》。发表于计算语言学协会年会（ACL），第 3639–3664 页，2025 年。[4] Nolan Bard、Jakob N. Foerster、Sarath Chandar、Neil Burch、Marc Lanctot、H. Francis Song、Emilio Parisotto、Vincent Dumoulin、Subhodeep Moitra、Edward Hughes、Iain Dunning、Shibl Mourad、Hugo Larochelle、Marc G. Belle mare 和 Michael Bowling。《Hanabi 挑战：人工智能研究的新前沿》。Artificial Intelligence，280：103216，2020 年。[5] Sushil Bikhchandani、David Hirshleifer 和 Ivo Welch。《作为信息级联的时尚、风潮、习俗与文化变迁理论》。政治经济学期刊，100(5)：992–1026，1992 年。[6] Bradley Brown、Jordan Juravsky、Ryan Ehrlich、Ronald Clark、Quoc V. Le、Christopher Ré 和 Azalia Mirhoseini。《大型语言猴子：通过重复采样扩展推理计算资源》。arXiv 预印本 arXiv:2407.21787，2024 年。[7] Mert Cemri、Melissa Z. Pan、Shuyi Yang、Lakshya A. Agrawal、Bhavya Chopra、Rishabh Tiwari、Kurt Keutzer、Aditya Parameswaran、Dan Klein、Kannan Ramchandran、Matei Zaharia、Joseph E. Gonzalez 和 Ion Stoica。《基于大语言模型的多智能体系统为何会失败？》。发表于神经信息处理系统进展（NeurIPS），数据集与基准赛道，2025 年。[8] Lingjiao Chen、Jared Quincy Davis、Boris Hanin、Peter Bailis、Ion Stoica、Matei Zaharia 和 James Zou。《增加大语言模型调用次数就是你所需要的一切吗？迈向复合人工智能系统的规模化特性》。发表于神经信息处理系统进展（NeurIPS），2024 年。[9] Weize Chen、Jiarui Yuan、Chen Qian、Cheng Yang、Zhiyuan Liu 和 Maosong Sun。《Optima：优化基于大语言模型的多智能体系统的有效性与效率》。发表于计算语言学协会论文集（ACL），2025 年。[10] Zijian Chen、Xueguang Ma、Shengyao Zhuang、Ping Nie、Kai Zou、Sahel Sharifymoghaddam、Andrew Liu、Joshua Green、Kshama Patel、Ruoxi Meng、Mingyi Su、Yanxi Li、Haoran Hong、Xinyu Shi、Xuye Liu、Hosna Oyarhoseini、Nandan Thakur、Crystina Zhang、Luyu Gao、Wenhu Chen 和 Jimmy Lin。《BrowseComp-Plus：用于深度搜索智能体的公平且解耦的评测基准》。发表于计算语言学协会年会（ACL），第 22349–22370 页，2026 年。[11] Hyeong Kyu Choi、Xiaojin Zhu 和 Sharon Li。《辩论还是投票：哪种方式能让多智能体大语言模型做出更好的决策？》。发表于神经信息处理系统进展（NeurIPS），2025 年。[12] Yufan Dang、Chen Qian、Xueheng Luo、Jingru Fan、Zihao Xie、Ruijie Shi、Weize Chen、Cheng Yang、Xiaoyin Che、Ye Tian、Xuantang Xiong、Lei Han、Zhiyuan Liu 和 Maosong Sun。《通过演化编排实现多智能体协作》。发表于神经信息处理系统进展（NeurIPS），2025 年。[13] Hanze Dong、Wei Xiong、Deepanshu Goyal、Yihan Zhang、Winnie Chow、Rui Pan、Shizhe Diao、Jipeng Zhang、Kashun Shum 和 Tong Zhang。《RAFT：用于生成式基础模型对齐的奖励排序微调》。机器学习研究汇刊，2023 年。[14] Yilun Du、Shuang Li、Antonio Torralba、Joshua B. Tenenbaum 和 Igor Mordatch。《通过多智能体辩论提升语言模型的事实准确性与推理能力》。发表于国际机器学习会议（ICML），2024 年。[15] Shangbin Feng、Yuyang Bai、Ziyuan Yang、Yike Wang、Zhaoxuan Tan、Jiajie Yan、Zhenyu Lei、Wenxuan Ding、Weijia Shi、Haojin Wang、Zhenting Qi、Yuru Jiang、Heng Wang、Chengsong Huang、Yu Fei、Jihan Yao、Yilun Du、Luke Zettlemoyer、Yejin Choi 和 Yulia Tsvetkov。《MoCo：模型协作研究的一站式平台》。arXiv 预印本 arXiv:2601.21257，2026 年。[16] Shangbin Feng、Yike Wang、Weijia Shi、Luke Zettlemoyer、Yejin Choi 和 Yulia Tsvetkov。《扩展模块化人工智能系统中的参与规模》。arXiv 预印本 arXiv:2606.07812，2026 年。[17] Chrisantha Fernando、Dylan Sunil Banarse、Henryk Michalewski、Simon Osindero 和 Tim Rocktäschel。《Promptbreeder：通过提示演化实现自指式自我改进》。发表于国际机器学习会议（ICML），第 13481–13544 页，2024 年。[18] Jakob Foerster、Gregory Farquhar、Triantafyllos Afouras、Nantas Nardelli 和 Shimon Whiteson。《反事实多智能体策略梯度》。发表于AAAI人工智能会议，第 2974–2982 页，2018 年。[19] GLM-4.5团队。《GLM-4.5：智能体化、推理与编码（ARC）基础模型》。arXiv 预印本 arXiv:2508.06471，2025 年。[20] Yuxian Gu、Li Dong、Furu Wei 和 Minlie Huang。《MiniLLM：大语言模型的知识蒸馏》。发表于国际学习表征会议（ICLR），2024 年。[21] Bochen Han 和 Songmao Zhang。《基于黑板架构探索先进的大语言模型多智能体系统》。arXiv 预印本 arXiv:2507.01701，2025 年。[22] Barbara Hayes-Roth。《用于控制的黑板架构》。Artificial Intelligence，26(3)：251–321，1985 年。

[23] Haoyang Hong、Jiajun Yin、Yuan Wang、Jingnan Liu、Zhe Chen、Ailing Yu、Ji Li、Zhiling Ye、Hansong Xiao、Yefei Chen、Hualei Zhou、Yun Yue、Minghui Yang、Chunxiao Guo、Junwei Liu、Peng Wei 和 Jinjie Gu。多智能体深度研究：使用 M-GRPO 训练多智能体系统。arXiv 预印本 arXiv:2511.13288，2025。[24] Lu Hong 和 Scott E. Page。由多样化问题求解者组成的群体能够胜过由高能力问题求解者组成的群体。《美国国家科学院院刊》，101(46):16385–16389，2004。[25] Jiaxin Huang、Shixiang Gu、Le Hou、Yuexin Wu、Xuezhi Wang、Hongkun Yu 和 Jiawei Han。大语言模型能够自我改进。载于自然语言处理实证方法会议（EMNLP），第 1051–1068 页，2023。[26] Jiedong Jiang、Wanyi He、Yuefeng Wang、Guoxiong Gao、Yongle Hu、Jingting Wang、Nailin Guan、Peihao Wu、Chunbo Dai、Liang Xiao 和 Bin Dong。FATE：面向多个难度等级前沿代数的形式化基准测试系列。arXiv 预印本 arXiv:2511.02872，2025。[27] Carlos E. Jimenez、John Yang、Alexander Wettig、Shunyu Yao、Kexin Pei、Ofir Press 和 Karthik Narasimhan。SWE-bench：语言模型能否解决真实世界的 GitHub 问题？载于国际学习表征会议（ICLR），2024。[28] Bowen Jin、Hansi Zeng、Zhenrui Yue、Jinsung Yoon、Sercan Arik、Dong Wang、Hamed Zamani 和 Jiawei Han。Search-R1：通过强化学习训练大语言模型进行推理并利用搜索引擎。载于语言建模会议（COLM），2025。[29] Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang 和 Ion Stoica。借助 PagedAttention 高效管理大语言模型服务中的内存。载于 ACM 操作系统原理研讨会（SOSP），2023。[30] Patrick R. Laughlin 和 Alan L. Ellis。数学智力任务中的可示范性与社会组合过程。《实验社会心理学杂志》，22(3):177–189，1986。[31] Junyou Li、Qin Zhang、Yangbin Yu、Qiang Fu 和 Deheng Ye。智能体越多越好。《机器学习研究汇刊》，2024。[32] Yiwei Li、Peiwen Yuan、Shaoxiong Feng、Boyuan Pan、Xinglin Wang、Bin Sun、Heda Wang 和 Kan Li。摆脱高昂成本：用于多步推理的提前停止式自洽性。载于国际学习表征会议（ICLR），2024。[33] Yuxuan Li、Aoi Naito 和 Hirokazu Shirado。多智能体大语言模型在分布式信息下集体推理的系统性失效。载于国际机器学习会议（ICML），2026。[34] Zhuofeng Li、Dongfu Jiang、Xueguang Ma、Haoxiang Zhang、Ping Nie、Yuyu Zhang、Kai Zou、Jianwen Xie、Yu Zhang 和 Wenhu Chen。OpenResearcher：用于长时程深度研究轨迹合成的全开放流程。arXiv 预印本 arXiv:2603.20278，2026。[35] Zhuofeng Li、Haoxiang Zhang、Seungju Han、Sheng Liu、Jianwen Xie、Yu Zhang、Yejin Choi、James Zou 和 Pan Lu。流内智能体化系统优化：实现有效规划与工具使用。载于国际学习表征会议（ICLR），2026。[36] Shuo Liu、Tianle Chen、Zeyu Liang、Xueguang Lyu 和 Christopher Amato。结合多智能体强化学习的大语言模型协作。arXiv 预印本 arXiv:2508.04652，2025。[37] Jan Lorenz、Heiko Rauhut、Frank Schweitzer 和 Dirk Helbing。社会影响如何损害群体智慧效应。《美国国家科学院院刊》，108(22):9020–9025，2011。[38] Irving Lorge 和 Herbert Solomon。尤里卡型问题求解中的两种群体行为模型。《心理测量学》，20(2):139–148，1955。[39] Yubo Ma、Yuhang Zang、Liangyu Chen、Meiqi Chen、Yizhu Jiao、Xinze Li、Xinyuan Lu、Ziyu Liu、Yan Ma、Xiaoyi Dong、Pan Zhang、Liangming Pan、Yu-Gang Jiang、Jiaqi Wang、Yixin Cao 和 Aixin Sun。MMLongBench-Doc：利用可视化对长上下文文档理解进行基准测试。载于神经信息处理系统进展（NeurIPS），数据集与基准赛道，2024。[40] Yuzhen Mao、Jerry Gu、Aadi Chauhan、Qizheng Zhang、Hangoo Kang 和 Azalia Mirhoseini。具有共享上下文的去中心化多智能体系统。arXiv 预印本 arXiv:2606.10662，2026。[41] Sumeet Ramesh Motwani、Chandler Smith、Rocktim Jyoti Das、Rafael Rafailov、Ivan Laptev、Philip H. S. Torr、Fabio Pizzati、Ronald Clark 和 Christian Schroeder de Witt。MALT：通过多智能体大语言模型训练提升推理能力。载于语言建模会议（COLM），2025。[42] OpenAI。gpt-oss-120b 和 gpt-oss-20b 模型卡片。arXiv 预印本 arXiv:2508.10925，2025。[43] Chanwoo Park、Seungju Han、Xingzhi Guo、Asuman Ozdaglar、Kaiqing Zhang 和 Joo-Kyung Kim。MAPoRL2：通过强化学习对协作式大语言模型进行多智能体后协同训练。载于计算语言学协会年会（ACL），第 30215–30248 页，2025。[44] Chen Qian、Zihao Xie、YiFei Wang、Wei Liu、Kunlun Zhu、Hanchen Xia、Yufan Dang、Zhuoyun Du、Weize Chen、Cheng Yang、Zhiyuan Liu 和 Maosong Sun。扩展基于大语言模型的多智能体协作。载于国际学习表征会议（ICLR），2025。[45] Qwen团队。Qwen3.5 模型。
https://modelscope.cn/models/Qwen/Qwen3.5-9B，2026。模型发布，2026年2月。

[46] Tabish Rashid、Mikayel Samvelyan、Christian Schroeder de Witt、Gregory Farquhar、Jakob Foerster 和 Shimon Whiteson。QMIX：用于深度多智能体强化学习的单调价值函数分解。载于国际机器学习会议（ICML），第 4295–4304 页，2018。[47] David Rein、Betty Li Hou、Asa Cooper Stickland、Jackson Petty、Richard Yuanzhe Pang、Julien Dirani、Julian Michael 和 Samuel R. Bowman。GPQA：研究生水平谷歌不可检索问答基准。载于语言建模会议（COLM），2024。[48] Jianhao Ruan、Zhihao Xu、Yiran Peng、Fashen Ren、Zhaoyang Yu、Xinbing Liang、Jinyu Xiang、Yongru Chen、Bang Liu、Chenglin Wu、Yuyu Luo 和 Jiayi Zhang。AOrchestra：自动创建子智能体以实现智能体编排。载于国际机器学习会议（ICML），2026。[49] Juergen Schmidhuber。哥德尔机器：实现可证明最优自我改进的自指通用问题求解器。arXiv 预印本 arXiv:cs/0309048，2003。[50] John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。近端策略优化算法。arXiv 预印本 arXiv:1707.06347，2017。[51] Zhihong Shao、Peiyi Wang、Qihao Zhu、Runxin Xu、Junxiao Song、Xiao Bi、Haowei Zhang、Mingchuan Zhang、Y. K. Li、Y. Wu 和 Daya Guo。DeepSeekMath：推动开放语言模型数学推理能力的极限。arXiv 预印本 arXiv:2402.03300，2024。[52] Lloyd S. Shapley。n人博弈的价值。载于 Harold W. Kuhn 和 Albert W. Tucker 编，《博弈论贡献》第二卷，数学研究年刊第 28 号，第 307–317 页。普林斯顿大学出版社，1953。[53] Shannon Zejiang Shen、Yapei Chang、Qiyao Liang、Benjamin Newman、Jenny Liang、Rulin Shao、Ziqiao Ma、Junjie Oscar Yin、Yung-Sung Chuang、Weijia Shi、Yejin Choi、Luke Zettlemoyer、Pang Wei Koh、Mohit Iyyer、Tongshuang Wu 和 David Sontag。World of SWECraft：通过模拟软件世界实现大规模智能体协作。arXiv 预印本 arXiv:2610.24061，2026。[54] Avi Singh、John D. Co-Reyes、Rishabh Agarwal、Ankesh Anand、Piyush Patil、Xavier Garcia、Peter J. Liu、James Harrison、Jaehoon Lee、Kelvin Xu、Aaron Parisi、Abhishek Kumar、Alex Alemi、Alex Rizkowsky、Azade Nova、Ben Adlam、Bernd Bohnet、Gamaleldin Elsayed、Hanie Sedghi、Igor Mordatch、Isabelle Simpson、Izzeddin Gur、Jasper Snoek、Jeffrey Pennington、Jiri Hron、Kathleen Kenealy、Kevin Swersky、Kshiteej Mahajan、Laura Culp、Lechao Xiao、Maxwell L. Bileschi、Noah Constant、Roman Novak、Rosanne Liu、Tris Warkentin、Yundi Qian、Yamini Bansal、Ethan Dyer、Behnam Neyshabur、Jascha Sohl-Dickstein 和 Noah Fiedel。超越人类数据：扩展基于语言模型的问题求解自训练。《机器学习研究汇刊》，2024。[55] Andries Petrus Smit、Nathan Grinsztajn、Paul Duckworth、Thomas D. Barrett 和 Arnu Pretorius。我们是否应该采用 MAD？浅析大语言模型的多智能体辩论策略。载于第 41 届国际机器学习会议（ICML）论文集，机器学习研究论文集第 235 卷，第 45883–45905 页，2024。[56] Garold Stasser 和 William Titus。群体决策中未共享信息的汇集：讨论期间的有偏信息采样。《人格与社会心理学杂志》，48(6):1467–1478，1985。[57] Ivan D. Steiner。群体过程与生产力。学术出版社，纽约，1972。[58] Vighnesh Subramaniam、Yilun Du、Joshua B. Tenenbaum、Antonio Torralba、Shuang Li 和 Igor Mordatch。多智能体微调：利用多样化推理链实现自我改进。载于国际学习表征会议（ICLR），2025。[59] Peter Sunehag、Guy Lever、Audrunas Gruslys、Wojciech Marian Czarnecki、Vinicius Zambaldi、Max Jaderberg、Marc Lanctot、Nicolas Sonnerat、Joel Z. Leibo、Karl Tuyls 和 Thore Graepel。基于团队奖励的合作多智能体学习价值分解网络。载于国际自主智能体与多智能体系统会议（AAMAS），第 2085–2087 页，2018。[60] Karthik Valmeekam、Matthew Marquez、Alberto Olmo、Sarath Sreedharan 和 Subbarao Kambhampati。PlanBench：用于评估大语言模型规划及变化推理能力的可扩展基准测试。载于神经信息处理系统进展，数据集与基准赛道，2023。[61] Jianhong Wang、Yuan Zhang、Tae-Kyun Kim 和 Yunjie Gu。Shapley Q值：解决全局奖励博弈的局部奖励方法。载于AAAI人工智能会议，2020。[62] Junlin Wang、Jue Wang、Ben Athiwaratkun、Ce Zhang 和 James Zou。智能体混合提升大语言模型能力。载于国际学习表征会议（ICLR），2025。[63] Qineng Wang、Zihao Wang、Ying Su、Hanghang Tong 和 Yangqiu Song。重新思考大语言模型推理的界限：多智能体讨论是关键吗？载于计算语言学协会第 62 届年会（ACL）论文集，第 6106–6131 页，2024。[64] Xuezhi Wang、Jason Wei、Dale Schuurmans、Quoc Le、Ed Chi、Sharan Narang、Aakanksha Chowdhery 和 Denny Zhou。自洽性提升语言模型中的思维链推理能力。载于国际学习表征会议（ICLR），2023。[65] David H. Wolpert 和 Kagan Tumer。集体成员的最优收益函数。《复杂系统进展》，4(2–3):265–279，2001。[66] Anita Williams Woolley、Christopher F. Chabris、Alex Pentland、Nada Hashmi 和 Thomas W. Malone。人类群体表现中存在集体智能因子的证据。《科学》，330(6004):686–688，2010。[67] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong

Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。《Qwen3 技术报告》。*arXiv* 预印本 *arXiv:2505.09388*，2025 年。[68] Qiying Yu、Zheng Zhang、Ruofei Zhu、Yufeng Yuan、Xiaochen Zuo、Yu Yue、Weinan Dai、Tiantian Fan、Gaohong Liu、Lingjun Liu、Xin Liu、Haibin Lin、Zhiqi Lin、Bole Ma、Guangming Sheng、Yuxuan Tong、Chi Zhang、Mofan Zhang、Wang Zhang、Hang Zhu、Jinhua Zhu、Jiaze Chen、Jiangjie Chen、Chengyi Wang、Hongli Yu、Yuxuan Song、Xiangpeng Wei、Hao Zhou、Jingjing Liu、Wei-Ying Ma、Ya-Qin Zhang、Lin Yan、Mu Qiao、Yonghui Wu 和 Mingxuan Wang。《DAPO：大规模开源大语言模型强化学习系统》。*arXiv* 预印本 *arXiv:2503.14476*，2025 年。[69] Weizhe Yuan、Richard Yuanzhe Pang、Kyunghyun Cho、Xian Li、Sainbayar Sukhbaatar、Jing Xu 和 Jason Weston。《自奖励语言模型》。载于国际机器学习会议（*ICML*），2024 年。[70] Eric Zelikman、Yuhuai Wu、Jesse Mu 和 Noah D. Goodman。《STaR：使用推理进行自举推理》。载于神经信息处理系统进展（*NeurIPS*），2022 年。[71] Eric Zelikman、Eliana Lorch、Lester Mackey 和 Adam Tauman Kalai。《自学优化器（STOP）：递归式自我改进代码生成》。载于语言建模会议（*COLM*），2024 年。[72] Guibin Zhang、Hejia Geng、Xiaohang Yu、Zhenfei Yin、Zaibin Zhang、Zelin Tan、Heng Zhou、Zhongzhi Li、Xiangyuan Xue、Yijiang Li、Yifan Zhou、Yang Chen、Chen Zhang、Yutao Fan、Zihu Wang、Songtao Huang、Francisco Piedrahita-Velez、Yue Liao、Hongru Wang、Mengyue Yang、Heng Ji、Jun Wang、Shuicheng Yan、Philip Torr 和 Lei Bai。《大语言模型智能体强化学习概况：综述》。机器学习研究汇刊，2025 年。[73] Hangfan Zhang、Zhiyao Cui、Jianhao Chen、Xinrun Wang、Qiaosheng Zhang、Zhen Wang、Dinghao Wu 和 Shuyue Hu。《停止过度重视多智能体辩论——我们必须重新思考评估并拥抱模型异质性》。*arXiv* 预印本 *arXiv:2502.08788*，2025 年。[74] Jenny Zhang、Shengran Hu、Cong Lu、Robert Tjarko Lange 和 Jeff Clune。《达尔文-哥德尔机器：自我改进智能体的开放式演化》。载于国际学习表征会议（*ICLR*），2026 年。[75] Wanjia Zhao、Mert Yuksekgonul、Shirley Wu 和 James Zou。《SiriuS：通过自举推理实现自我改进多智能体系统》。载于神经信息处理系统进展（*NeurIPS*），2025 年。[76] Yujie Zhao、Lanxiang Hu、Yang Wang、Minmin Hou、Hao Zhang、Ke Ding 和 Jishen Zhao。《Stronger-MAS：面向协作式大语言模型的多智能体强化学习》。载于国际学习表征会议（*ICLR*），2026 年。[77] Yuxin Zuo、Kaiyan Zhang、Li Sheng、Shang Qu、Ganqu Cui、Xuekai Zhu、Haozhan Li、Yuchen Zhang、Xinwei Long、Ermo Hua、Biqing Qi、Youbang Sun、Zhiyuan Ma、Lifan Yuan、Ning Ding 和 Bowen Zhou。《TTRL：测试时强化学习》。载于神经信息处理系统进展（*NeurIPS*），2025 年。

A. 附录概述

附录 B 证明了第 4 节中的命题。附录 C 规定了等预算阶梯，附录D 介绍了隐藏档案协议、稀缺注意力和结构读取器。附录 E描述了令牌核算，附录 F介绍了递归实验，附录 G说明了将某项结果命名为递归的规则。附录H 描述了黑板和智能体测试框架，附录 I 介绍了构造任务，附录 J 介绍了六个共享信息基准测试的评估协议。附录 K 报告了其他结果，包括信用实验的设置；附录L 介绍了在 PlanBench、HiddenBench 和 Hanabi 上进行递归训练，附录M介绍了计算资源和复现。

B. 证明

命题 1。（i）一组 n 个独立成员只有在这 n 个成员全部失败时，才会在问题 x 上失败；其发生概率为 $(1-p_x)^n$，因此 $T(n) = 1-\mathbb{E}_x[(1-p_x)^n]$。增量 $T(n+1) - T(n) = \mathbb{E}_x[p_x(1-p_x)^n]$ 非负，且关于 n 单调不增，因此 T 单调不减且为凹函数。我们的参照协议从一个包含 m 个成员的样本池中无放回地抽取成员，其中有 k_x 个成员在问题 x 上答对。此时失败概率为 $q_n = \prod_{j<n}(m-k_x-j)/(m-j)$，增量 $q_n - q_{n+1} = q_n k_x/(m-n)$ 为非负，并且连续增量之比为 $(m-k_x-n)/(m-n-1) \le 1$（当 $k_x \ge 1$ 时），因此相同结论仍然成立；网格规模之间的线性插值会保留这些性质。（ii）设某规则观察答案 a_1、……、a_n，并承诺给出某个 a_j。对于任何问题，承诺答案只有在至少有一个 a_i 答对时才可能正确，因此该规则在该问题上的准确率至多等于至少有一个成员答对的概率；对问题和抽样取平均后，准确率至多为 $T(n)$。该论证适用于任何选择 j 的方式，因此涵盖投票、置信度加权、验证器和评判器。对 n 个成员进行多数投票就是这样一种规则，因此 $V(n) \le T(n)$，并且根据第（i）部分，$V(1) = T(1) \le T(n)$。因此，超过 $T(n_B)$ 的群体也超过 $V(1)$，并且同样超过 $V(n_B)$。（iii）根据第（i）部分，T 的增量从不超过第一个增量，即 $T(1) = V(1)$，因此 $T(n) \le n\,V(1)$，而 $T(n) \le 1$ 显然成立；对于无放回抽样，第一个增量为 k_x/m，同样等于一个成员的准确率。相同的上界给出：对于 $n > m$，$T(n) - T(m) \le (n-m)\,V(1)$。□

命题 2。设群体报告了 r 个不同声明，其中 h 个为真实声明，并有 t 个真实发现；再加入一个校准置信度为 q（即该声明为真的概率）的声明。对于一个关于真实声明和虚假声明数量呈线性的评分，该声明的预期影响等于 q 倍真实声明的影响加上 $(1-q)$ 倍虚假声明的影响。对于真实声明，召回率 h/t 增加 $1/t$，对于虚假声明则保持不变，因此预期变化 q/t 对每个 $q > 0$ 都为正。F1 值等于 $2h/(r+t)$。加入该声明后，若声明为真，则 F1 值等于 $2(h+1)/(r+1+t)$；否则等于 $2h/(r+1+t)$，因此其期望值为 $2(h+q)/(r+t+1)$。当且仅当 $(h+q)(r+t) > h(r+t+1)$ 时，该值超过 $2h/(r+t)$，也就是当 $q > h/(r+t) = \mathrm{F1}/2$ 时。置信度门控声明的得分为 $(h - c(r-h))/t$。真实声明增加 $1/t$，虚假声明增加 $-c/t$，因此预期变化 $(q - c(1-q))/t$ 当且仅当 $q > c/(1+c)$ 时为正。□

置信度门控声明的 Shapley 份额。设 $P_S = \bigcup_{i\in S} P_i$，并令 $w(p) = +1$ 表示真实声明，$-c$ 表示虚假声明。于是 $v(S) = \frac{1}{|T|}\sum_{p\in P_S} w(p) = \frac{1}{|T|}\sum_p w(p)\,u_p(S)$，其中 $u_p(S) = \mathbb{1}[\exists i \in S:\ p \in P_i]$。每个 u_p 都是一个博弈：只要联盟包含报告了 p的 k_p 成员之一，该联盟就获胜。根据对称性和空参与者性质，每位报告者的 Shapley 份额为 $1/k_p$，其他所有人的份额均为 0。根据可加性，$\phi_i = \frac{1}{|T|}\sum_{p\in P_i} w(p)/k_p$，因此每项声明的收益或成本由其报告者平均分配，且份额总和为 R。□

平均信用额度。 将第 4.3 节的信用额度对成员求和，可得 $N\lambda R + (1-\lambda)N\sum_i f_i$。流量在成员之间转移正份额，并保留其收到的每个单位，因此 $\sum_i f_i = \sum_i \phi_i = R$，平均信用额度为 R。□

C. 等预算阶梯上的评分

单位成本。 c_1 是相同模型、协议族和平台的独立池中完成令牌总数除以作答的样本池成员数。因而，从未作答的成员虽然获得了报酬，却不计入人数。DocSweep 的问题成本各不相同，因此采用按问题计算的 $c_1(x)$。群组记录的预算是其完成令牌数除以 c_1，而单元格的预算 n_B 是其中各条记录预算的平均值，因此 n_B 当某个协议的支出低于独立成员的平均支出时，该值会小于 1。我们报告的每个层级均以完成令牌为单位。

引用。 V（n）是将群体自身的承诺规则应用于从样本池中无放回抽取的 n 名成员后得到的期望得分。该规则选择支持者最多的答案；若出现平局，则选择明确置信度总和较高的答案；若仍然平局，则将剩余结果平均分配。 T（n）是 n 名成员中至少有一名答对的概率。两条曲线均根据网格 1、2、3、4、6、8、12、16、24、32、48个成员中每个问题的正确答案和错误答案数量计算得出

名成员之间进行线性插值；DocSweep 会在每个整数规模上精确计算这些值。在隐藏配置文件任务中，抽样按席位进行分层，因此大小为 n 的集成包含的席位与真实群体的比例相同，而群体规模倍数之间的规模则尽可能均匀地抽取。分层 T 是精确的，而 V 则通过枚举精确计算，或根据每个问题和规模至少 2,000 次抽样进行估计。严格样本池绝不会读取超出其自身规模的参照协议。预算超过其样本池最大规模的单元格，会在第 1 级和第 2 级按照该规模的参照协议进行评分；其第 3 级在该单元中记录为不可评分，显示为已达到的级别并标记为已封顶。

样本池大小。 从预算上限实验开始，一次小规模试运行会在主要单元运行前测量每个单元的预算。随后，样本池为每个问题提供 $S\lceil 1.5$ 最多 $n_B/S\rceil$ 名成员，用于由 S 名成员组成的群体，且至少有64名成员；四名成员的预算上限实验为每个问题分配了68名成员。表2中的HiddenBench阶梯使用较早的每项任务32名成员的样本池，而DocSweep的样本池为每个问题16名成员。

检验。 每个级别都是显著性水平为0.025的单侧检验，用于检验 Y 与参照协议之间的配对差值 k，并采用按问题聚类的标准误；各级别按顺序检验：当该检验确认级别1至 k时，一个单元达到级别 q。未经确认的正边际会使该单元停留在低一级，而低于一名成员的单元处于级别0。在一项任务内，针对该任务各单元执行Benjamini–Hochberg程序；当其任务内 q 超过0.025时，将通过标记为不稳健，但不会改变级别。与投票的一致性通过固定边际下的两个单侧检验进行检验。

A/A检验。 每项实验都指定自己的A/A检查。样本池可以按随机种子或成员位置分成两半，用一半的集成对照另一半的参照答案进行评估；也可以用样本池运行构成的伪群体对照不包含这些运行的样本池进行评估；或者用同一批次的新随机种子对照现有样本池进行评估。这些检查在投票和Pass@n 层级运行，界限设定为包含零的95%区间与 $|z| \le 2.58$之间。

D. 隐藏档案协议与稀缺注意力

方法名称。 表 8 按成员交换的内容对协议进行分类，表格使用以下简称。

- **独立参考答案** (独立)。单一 表示一个独立成员， V(1)。*Vote@B* 和 *Pass@B* 分别表示对成本相同的独立成员进行多数投票和Pass@n， V(n_B)和 T(n_B)。*Sequential-4* 表示一个智能体依次作答四次。
- **交互基线** (辩论)。黑板 是黑板，而 *Debate-2R* 和*Debate-3R* 分别是进行两轮和三轮的辩论 [14]。*Self-Refine-2R* 是两轮自我纠正，其中每个成员仅读取自己的文本；后缀+*H* 表示添加了提示，指出事实可能是分布式的。
- **证据协议**（证据）。事实转发让成员只能转发各自获得的事实，而共享上下文则附上每位成员的事实及自由评论；*SC*+转发结合了二者。仅事实 *BB*是只接受转发事实的黑板，而门控*BB*则是只有完整读取者才能得出结论的黑板。
- **稀缺注意力**。无传输为每位成员提供一个读取单位，自组织允许预算转移，并且 +规划轮次增加一个两步骤的规划轮次。预算上限增加一个上限，使单一枢纽不可行；无中继禁止在该上限下进行中继（全部属于涌现）；自编写提示词则根据规则编写，或经过多代演化而成。枢纽和树是设计的控制方案（设计式），而遵循自编写提示词的群体，无论是否设有上限，均属于自提示。
- **DocSweep**。自组织（涌现）、自编写提示词（自提示）和聚合器（设计式）分别是由四名成员组成的群体：它们分别进行自组织、遵循自编写提示词，或遵循设计的聚合器。
- **递归**(自训练)。*G0* 是未训练模型，*G1* 则是基于其经协议核验的首轮发言训练而成的第一代。*G2* 是第二代，基于从 G1 中抽取的每个席位的最短经核验首轮发言训练而成；对照组则是它的对照组，使用相同算子，从同一父代抽取 G0 的首轮发言进行训练。

任务。 每个 HiddenBench 任务都是在三个选项中做出决策。共享事实支持一个诱饵选项，而私有事实合起来支持另一个选项 [33]。其中 58 个任务恰好包含四条私有事实。我们排除了一项其事实为表格行的任务。随后进行两项离线检查：如果某条私有事实包含另一条事实的关键词，或合并后的消息会被认定包含某条缺失事实，则移除相应任务，最终剩下 56 个任务。生成的任务遵循相同模板：由 Qwen3.8-27B 编写，并指定一个诱饵选项和一个答案；当获得全部事实的成员意见一致、仅获得共享事实的成员受到误导，且仅获得各自事实的成员大多失败时，该任务才会被纳入。编写者的答案键不会保存在任务文件中，仅用于报告标签质量。

表8：我们纳入阶梯的协议，按成员交换的内容以及组织方式由谁决定分组，并标注图5a中的家族代码。附录 D 定义了各个简称。

家族	交换	组织方式由谁决定	协议
独立 独立参考答案	无	none	单一、Vote@B、Pass@B、Sequential-4
辩论 论证交换	答案、理由	固定轮次	黑板、Debate-2R、Debate-3R、Self-Refine-2R、 +H
证据 证据交换	给定的事实	固定协议	事实转发、共享上下文、SC+转发、仅事实 BB、门控BB
涌现自建组织	阅读预算，中继节点	成员	无传输、自组织、 + 规划轮次、预算上限、无中继
自提示书面组织	自编写提示词	模型	从规则演化而来，各自分为有预算上限和无预算上限两种
设计式设计式组织方式	指定角色	设计者	枢纽、树、聚合器
自训练训练过的成员	经验证的交互轮次	训练	G0、G1、G2、对照组

消息与预算。 成员将消息发布到黑板上，其中包含子任务标题和正文。转发工具会逐字复制事实，主机会根据任务中的事实检查每一行转发内容，并拒绝并非事实的行。未读取消息的标题会说明该消息包含谁的私有事实，具体由主机根据消息内容计算得出。成员打开另一名成员的消息时，会消耗阅读预算。转移会将预算单位转给另一名成员；如果转移会超过上限，主机会拒绝该转移并说明上限。只有持有全部私有事实的成员才能发布分析，而分析会推送给所有成员。时序规则规定，在第 16 步之前暂缓提交，除非已有分析到达。

设计的控制方案与自写提示词。设计型中心获得全部阅读预算。面向四名成员的设计树为中心分配两个单位，为中继节点分配一个单位；两步后开放中继节点的读取权限，四步后开放中心的读取权限，并为每名成员提供角色卡，说明应读取谁的消息以及应转发什么内容。模型根据一段说明协议规则的文本编写提示词。演化提示词经过四代修订，每代包含 16 项生成任务；反馈描述群体的流程、转移、中继和拒绝情况，而绝不涉及答案。

结构读取器。对于每条记录，读取器都会找出最先掌握每项私人事实的成员，且不依赖被推送的分析。它掌握的每项事实都有一个深度：所有者转发的事实深度为一，中继事实的深度则比其源消息多一。若从读取器的读取集合中移除某条消息后，读取器在完成的步骤中便无法掌握所有事实，则该消息是必要的。如果读取器仅通过深度为一的消息完成读取，则该记录为星形；若最深的必要消息深度为二，则为两层树；若深度为三或更高，则为深层中继。空策略保留其预算或随机分配预算，前两步转发自身事实，之后随机均匀地读取；我们在每种协议下针对每个问题模拟该策略 10,000 次。

E. 协调令牌核算

每个成员轮次都与其模型调用相匹配，并根据其工具调用进行分类。当某个轮次未进行工具调用、仅等待或持有提交内容时，我们将其计为等待。转移，以及成员开始读取前发布或读取的内容，均计为协商。其余所有轮次均为工作。完成令牌和新读取的提示词令牌按类别求和。不含等待的重放会移除所有等待轮次，并保留其他轮次的记录，同时固定准确率，因此它给出了这样一种协议节省量的上界：该协议让等待中的成员处于空闲状态。通过从每个轮次的完成令牌中减去可见输出的上界，可以得到等待轮次内部思考量的下界。令牌核算仅用于描述，其逐单元检验未经校正。

F. 递归实验

设置。 第0代是基础模型。第1代在其父代群体于 SplitBench 训练任务上的首轮发言上进行训练：这些任务会写出成员的每个私有事实，且不包含立场；使用两个训练随机种子，从而形成两个谱系。每个谱系的第二代从自身的第1代中为每个席位采样四个首轮发言，并在其中符合标准的最短发言上进行训练。训练采用全参数监督微调，批次大小为8，学习率为2×10^{-6}。三步辩论和27B的第1代与第二代仅使用经核验的首轮发言训练240步。9B标准辩论运行的第二代训练320步，并加入承诺轮次，即仅训练成员提交调用的骨架、其损失权重为四分之一；4B运行使用其自身的算子（见下文）。每一代都在196个SplitBench留出任务上进行评估，参照对象是由基础模型构成的冻结池，每个问题包含32个独立成员；此外，还在65个HiddenBench任务这一次要集合上进行评估。对比

表9：准确率、成本和协调开销。（a）在 HiddenBench 上将一个协议与参考协议进行对比，两者均在同一样本池上评估：其预算为 n_B，该协议相对于 pass@n 的边际为 $Y - T$（n_B），参考协议的边际（参考），两者之差拆分为准确率项和成本项，即边际 $-$ 参考协议 $=$ 准确率 $+$ 成本（公式1）；还包括该协议在参考预算下的边际（参考预算下），以及支撑其第3级的因素（经由）：当参考预算下的边际经确认大于零时为准确率；当仅在自身预算下的边际为正时为准确率与成本的共同作用；当两者均不满足时则没有因素。Debate-2R+H 是带提示的两轮辩论；Self-organized（56个任务）是在这56个任务上采用自由预算的自组织群体（准确率为0.690，n_B 23.4），与表3中的有上限群体属于不同单元格；第1–7行和第11行使用65个任务，其余各行使用包含四个私有事实的56个任务；#：图9中的数据点。**（b）**56个任务上的令牌核算：等待所耗令牌的占比（等待），以及所有协调所耗令牌的占比（协调）；同时给出当前预算和边际，以及移除每个等待回合后的预算和边际（自由等待）。 ±：标准误。

(a)

#	协议	参照协议	预算	边际	Ref.	准确率	Cost	参考预算下	Via
证据 协议									
1	事实转发，9B	Debate-2R+H	9.5/7.2	$+0.160 \pm 0.057$	-0.345	$+0.550 \pm 0.039$	-0.045 ± 0.005	$+0.205 \pm 0.057$	准确率
2	事实转发，27B	Debate-2R+H	6.2/8.8	$+0.310 \pm 0.049$	$+0.079$	$+0.177 \pm 0.033$	$+0.054 \pm 0.007$	$+0.256 \pm 0.049$	准确率
3	事实转发，35B	Debate-2R+H	10.6/9.5	$+0.181 \pm 0.049$	-0.176	$+0.373 \pm 0.036$	-0.016 ± 0.002	$+0.197 \pm 0.049$	准确率
4	共享上下文 ，27B	Debate-2R+H	7.9/8.8	$+0.190 \pm 0.048$	$+0.079$	$+0.099 \pm 0.025$	$+0.012 \pm 0.002$	$+0.178 \pm 0.048$	准确率
5	SC+转发，27B	共享上下文	6.0/7.9	$+0.328 \pm 0.049$	$+0.190$	$+0.091 \pm 0.023$	$+0.047 \pm 0.006$	$+0.282 \pm 0.050$	准确率
设计式与 自提示组织方式，*27B*									
6	设计型枢纽	自组织	11.3/21.3	$+0.130 \pm 0.056$	-0.124	$+0.187 \pm 0.034$	$+0.067 \pm 0.014$	$+0.063 \pm 0.055$	准确率+成本
7	设计型中心	自组织 + 规划	11.3/18.3	$+0.130 \pm 0.056$	$+0.037$	$+0.041 \pm 0.028$	$+0.052 \pm 0.010$	$+0.078 \pm 0.055$	准确率+成本
8	自编写提示词（规则）	自组织，56 个任务	9.2/23.4	$+0.205 \pm 0.055$	-0.100	$+0.190 \pm 0.036$	$+0.115 \pm 0.019$	$+0.090 \pm 0.050$	准确率+成本
9	自编写提示词（演化）	自组织，56 个任务	9.7/23.4	$+0.180 \pm 0.055$	-0.100	$+0.172 \pm 0.037$	$+0.109 \pm 0.018$	$+0.072 \pm 0.050$	准确率+成本
10	设计树	自组织 + 规划，上限	10.8/44.4	$+0.107 \pm 0.051$	-0.180	$+0.125 \pm 0.035$	$+0.162 \pm 0.030$	-0.055 ± 0.041	准确率+成本
涌现的 组织方式和预算受上限限制的 自提示，*27B*									
11	自组织 + 规划	自组织	18.3/21.3	$+0.037 \pm 0.047$	-0.124	$+0.146 \pm 0.033$	$+0.015 \pm 0.004$	$+0.021 \pm 0.047$	无
12	上限 + 提示词（演化）	自组织 + 规划，上限	25.0/44.4	$+0.043 \pm 0.041$	-0.180	$+0.163 \pm 0.027$	$+0.060 \pm 0.014$	-0.017 ± 0.039	无
13	自组织 + 规划，上限	无中继，上限	44.4/37.0	-0.180 ± 0.046	-0.267	$+0.106 \pm 0.025$	-0.019 ± 0.005	-0.161 ± 0.045	无

(b)

	共享		预算		ΔPass@n	
方法	Wait	协调	现在自由等待		Now	自由等待
自组织 + 规划，56 项任务	0.258	0.67	22.7	16.9	$+0.066$	$+0.100 \pm 0.047$
自组织 + 规划，上限	0.377	0.69	44.4	27.7	-0.180	-0.131 ± 0.045
上限 +提示词（规则）	0.404	0.67	40.2	23.9	-0.115	-0.061 ± 0.047
上限 +提示词（演化）	0.332	0.64	25.0	16.7	$+0.043$	$+0.090 \pm 0.044$
设计树，上限	0.308	0.50	10.8	7.5	$+0.107$	$+0.155 \pm 0.052$

按问题配对，并通过对问题进行 2,000 次自助法抽样来计算标准误，每次抽样都会重新计算 n_B。在一个谱系中，首轮发言涵盖所有私人事实的成员占比在第一代中增至三倍。

三步辩论中的 9B。 成员最多进行三个步骤，随后仅被要求回答一次。若对比超过 1.96 个标准误且预算未超出样本池，则该对比成立；对照组对比使用附录 G中的乘数。第二个步骤会将 $+0.187$ 加到边际上，并将成员的首次调用减半。对照组使用相同的父代、步骤和随机种子，重复第二代的算子，但从基础模型中抽取其首轮发言。第二代比 $+0.192$ 高，因此三项对比均成立，命名规则将该结果称为 9B 的权重级递归。两个步骤在 HiddenBench 上也都成立。

跨训练随机种子的复现。 第二代及其对照组使用相同的父代和行文件，在另外四个训练种子上重新训练。合并这些结果后，对比为 $+0.115 \pm 0.008$，且在全部八组配对中均为正；即使误差包含训练随机性，结果也依然如此，而准确率贡献了其中的大部分。

三步辩论中的 27B。 该 9B 方案在 27B 上重复进行。第一代模型以基准群体经核验的首轮发言进行训练，并以相同问题和席位的标准辩论前缀作为上下文，使边际从 -0.009提升至 $+0.201$；即使在 HiddenBench 上，以及当误差包含训练随机性时，这一步仍然成立。第二代模型使用第一代模型经核验的交互轮次中的标准辩论行进行训练，在较低预算下将边际提升至 $+0.284$，并且 $+0.268$ 高于使用相同方案在基础模型轮次上训练的对照组。表4列出了这些边际。

4B 运行。 4B 运行沿用标准辩论中的设计，并配有专属算子：它保留每个席位最短且经验证的首轮，并加入承诺轮次，其中只有提交调用的骨架会接受训练。第一步将边际提高 $+0.421$，在冻结基础池和 HiddenBench 上均保持这一结果，并且在每个

模型自身的样本池中也保持这一方向。第二代加入 $+0.194$，并达到第三个层级。

标准辩论中的 9B。 相同的 9B 设计在标准辩论中带来两步增益，分别为 $+0.112$ 和 $+0.179$（评估种子为 10 至 13）。在种子 14 至 17 上，表 4 报告的两步增益分别为 $+0.138$ 和 $+0.188$；第二步的每条谱系多使用了四个训练种子，且两步都提升了 65 个官方 HiddenBench 任务上的边际。第二代比一个使用 G0 回合进行同样训练的对照组高 $+0.116$，但由于测试框架更常关闭对照组的成员，命名规则使这一对比无法定论。

增益来自群体。 群体的表现可能有所提升，是因为其成员学会了解决任务，也可能是因为他们学会了汇集各自掌握的信息。在标准辩论运行中，每个模型也会在 65 个 HiddenBench 任务中的每个任务上，使用全部事实、以单独方式在八个随机种子上作答。第一步中，单独准确率的变化保持在 0.03 以内；而在成员各自掌握分散事实、且两代成员都能单独解决的任务上，群体准确率提高了 $+0.177$；第二步又增加了 $+0.064$（表 10a）。三步辩论的每一代也会与由该代成员组成、独立工作的样本池进行评估（表 10b）。在 9B 下，两步在各代自己的样本池对照下均成立；第二代的表现仍高于其成员的 pass@n，而单个成员单独工作时几乎没有提升，因此 9B 下的两步都属于集体增益。在 27B 下，第一代的表现远高于其自身成员，而训练后的成员单独工作时并不优于基础模型成员，因此第一步属于集体增益。第二代的表现仍高于其自身成员；它在冻结基础池上的这一步保留了父代在较低预算下的准确率，因此根据公式（1），这一步的增益来自成本项。

监督者重写27B团队。一位使用成员权重的监督者，在 SplitBench 训练任务上读取 27B 第一代群体的无标签记录，并逐个版本重写团队角色与指令，全程不使用答案键。在 196 项留出任务上，最终重写的团队 $+0.111$ 优于固定团队，且 $+0.210$优于使用相同权重、仅依据规则编写的团队。

G. 递归命名规则

命名规则规定了递归实验结果的命名方式。当等预算边际上的三项对比均成立时，结果才称为递归：第一代优于基础模型，第二代优于其父代，而第二代还优于一个对照组；该对照组与第二代拥有相同的父代、步数、采样预算和验证，其训练回合则取自父代的上一代。三个模型规模的递归端点在单侧显著性水平 0.025 下构成一个家族，并通过 Holm逐步下降法进行检验；该方法使用的乘数 2.39、2.24 和 1.96 分别是显著性水平为 0.025/3、0.025/2 和 0.025 时的正态分位数。9B 对照组使用乘数 2.24 进行评判。其比率为 12.0，高于该家族最大的乘数 2.39，因此无论排序位于何处，均可通过检验。使用另外四个随机种子进行的复现采用 2.50，其他所有端点均采用 1.96。

H. 共享黑板与智能体测试框架

黑板。 黑板是一个只追加的消息序列。一条消息包含四个字段：它所处理的子任务，措辞应采用后续搜索者会使用的表达方式；结果；简短的理由；以及它对较早消息或任务资源（例如代码片段、文档、文件和 Mathlib声明）的引用。消息写入时会获得下一个地址，且永不编辑或删除，因此引用构成一个有向无环图，该图只在最新端增长。子任务与结果合计在审计和检索中不得超过 256 个令牌，在证明和代码中不得超过 1,024 个令牌；一条消息最多包含 32 个引用。只有在每个引用都存在、且理由点名每个引用并说明其贡献时，发布才会被接受；否则不会写入任何内容，成员会得知哪个引用失败。

搜索与排序。 搜索 接收自由文本并返回八条最佳消息，而 访问 则按地址读取最多八条消息。两者都会返回四个字段、该消息所引用的消息，以及最多八条引用它的最新消息。作者和排序权重保持隐藏。每条消息都带有一个权重 $\omega_m \in [0,1)$，初始值为零；每当有新消息引用它时，权重就会增长：$\omega_m \leftarrow \omega_m + 0.3(1-\omega_m)$ 并在 $\omega_m \leftarrow 0.98\,\omega_m$ 每个群体步骤中衰减。查询 q会根据具有正 BM25 相关性的消息对其进行排序，排序依据为 $\omega_m + \mathrm{Rel}(q,m)/\max_{m'} \mathrm{Rel}(q,m')$；若出现并列，则优先较早的消息。因此，权重会重新排列相关消息，并排除不相关的消息。

测试框架。 所有成员接收相同的提示词和工具，并在全局步进时钟上异步运行：在每个步骤中，每个活跃成员轮流执行一次动作，即进行一次模型调用及其工具调用。每次工具调用均计为一次动作，

表10：群体层面的增益。（a）在65项 HiddenBench 任务上，9B 模型中单个成员在获得全部事实时的准确率（单独）、其变化（个体），以及在分散事实条件下、针对两代成员都能独立解决的任务，群体准确率的变化（团队）。**（b）**三步辩论：每一代分别以基础模型成员的 pass@n（基础 pass@n）或其自身成员独立运行的 pass@n（自身 pass@n）为基准进行评估；每个谱系使用四个训练随机种子运行9B；未标注误差的数值为点值。每一行9B数据都是基于自身记录的独立测量，因此两项增益之和不一定等于 G2 相对于其自身 pass@n 的边际。

（a）步骤	Solo	个体	Team
G0 →G1	0.927 → 0.926	−0.001 ±0.009	+0.177 ±0.038
G1 →G2	0.926 → 0.909	−0.017 ±0.008	+0.064 ±0.028

（b）读取	值
Qwen3.5-9B	
G0 相对于基准 pass@n	−0.253 ±0.027
G2 相较自身 pass@n	+0.082 ±0.025
G1 相较 G0 的增益，自身 pass@n	+0.197 ±0.026
G2 相较 G1 的增益，自身 pass@n	+0.127 ±0.017
单独成员 , G2 减去 G1	+0.005 ±0.003
Qwen3.8-27B	
G1 相对于自身 pass@n	+0.238 ±0.032
G1 相对于 G0 的增益，自身 pass@n	+0.245 ±0.027
G1 自身 pass@n 减去基准 pass@n	−0.037
G2 相对于自身 pass@n	+0.234 ±0.032

包括被拒绝的成员。当某个成员达到其动作上限，或运行达到其步骤上限时，该成员会从截至目前的工作中提交其最佳答案。审计、检索、证明和文档成员的历史记录在达到十个回合后，由同一模型进行总结，并保留最近三个回合；代码修复成员则保留完整历史记录。采样使用温度 1 和原生工具调用。在评估期间，一次调用最多生成 16,384 个令牌，在训练期间最多生成 8,192 个令牌；但开场的无工具思考可使用 32,768 个令牌。推理计算是所有调用的提示词令牌与完成令牌之和，包括摘要。表格11列出了这些限制。

I. SplitBench、矛盾审计与 DocSweep

SplitBench。 SplitBench 遵循 HiddenBench 的格式：共享事实会指向一个诱饵，而推翻这一结论的事实则分散在各成员之中。通义千问 Qwen3.8-27B 编写了 1,500 个候选任务，其中没有任何一个在措辞上与 HiddenBench 任务接近（词集 Jaccard 相似度最高为 0.143）。筛选仅依据同一模型的行为，每种条件下设置四个独立成员。如果成员不汇集各自的事实就能解出任务、仅凭共享事实无法使成员指向诱饵，或获知全部事实的成员意见不一致，则剔除该任务。通过筛选的 697 个任务按其标识符的哈希值划分为 501 个训练任务和 196 个留出任务；在开始任何一代的训练之前，这两个集合均已冻结。在通过筛选的任务中，获知全部事实的成员所达成的共识，有 98.6% 与编写者的答案键一致。

语料库。 评估语料库为 16 个项目的 Markdown 和 reStructuredText 文档，均从指定的提交获取：CPython、Docker、Godot、JAX、Kubernetes、Linux 内核、LLVM、NumPy、pandas、PyTorch、Ray、SciPy、scikit-learn、SQLAlchemy、SymPy 和 Transformers。我们保留参照协议页面和指南页面，排除介绍版本发布、计划、模板或许可的页面。我们采用贪心方式将段落打包成约 1,024 个令牌的代码片段，不拆分段落，并在每个代码片段的首行标明其文件名。训练采用相同的方法处理另外 35 个项目（共 21,430 个代码片段），其中 4 个项目留作验证。

植入矛盾。 候选片段对至少共享两个罕见词。模型最多列出三个共同主题，并改写其中一个片段的自然语言文本，使其与另一个片段相矛盾，同时不改动代码、表格、列表项和标题。机械检查将编辑范围限制为不超过文本的 30% 且不超过八处改动，要求改动不只是格式调整，并要求每个改动后的句子在语料库中都是唯一的。四位独立评判器必须一致认可这组片段自然、并非无关紧要、前后一致且确实存在冲突。每个语料库都包含 10 组同文件矛盾和 10 组跨文件矛盾，涉及 40 个不同的代码片段。该级别 $\times n$ 添加 $40(n-1)$ 个使用固定随机种子抽取的干扰片段，优先选取完整文件并保留原有顺序；我们在 $\times 4$ 处进行评估。

工具、答案和分数。成员通过以下方式读取代码片段：`fetch_snippet`，并使用以下方式搜索它们：`search_snippet`，这是一种精确且不区分大小写搜索，最多返回十个匹配项及其上下文。声称的矛盾是一对代码片段

表11：每个成员的　　测试框架限制（评估；训练在审计上使用60个操作和步数，在检索上使用80个操作和步数）　　.

	审计	检索	证明	代码修复	PDF文档和长文档	多项选择
每位成员的操作数	250	300	100	300	100	16
每次运行的步数	250	400	120	400	120	24
摘要	yes	yes	yes	no	yes	no

其理由逐字引用了两个代码片段中的冲突文本。群体的交付成果是其成员各不相同的配对的并集。精确召回率统计找到的植入配对中两个代码片段均正确的配对数，精确率是已提交配对中属于植入配对的比例，而精确配对 F1 值则结合了这两者。基准测试的原始分数在任何报告的配对包含其经过编辑的代码片段时，都会将其计为一个矛盾；由于这会奖励提交很长的配对列表，因此我们仅出于可比性报告该分数。在分布式设置中，代码片段会被随机分成 N 个份额，每个成员只能打开自己的份额，因此成员只能通过基于他人发布的内容，找到跨份额的矛盾；任务提示词对所有方法都如此说明。

DocSweep。 128 个问题中的每一个都需要结合可搜索语料库中散落在多份文档里的事实。一类问题要求跨文档整合事实，另一类则要求汇总多份文档中的信息。若读者打开所有相关文档后能够回答问题，且共享语料库没有提供捷径，该问题才会被纳入。此外，纳入的问题还必须满足：单个受上限约束的成员在四个随机种子中至多答对一个。由于纳入标准依据的是 27B 模型自身的行为，这组问题是为该模型筛选的，其外部效度有限。每位群体成员最多可执行 12 次操作，因此没有任何单个成员能打开所有相关文档。自行组织的成员会在共享看板上认领文档，在读取前的一个时间窗口内制定计划，并遵循时序规则。设计的聚合器会分配语料库份额，并整合成员发布的摘要；模型则根据规则编写提示词，一次定稿后不再修改。

J. 评估协议与基线

BrowseComp-Plus。 我们在包含 100,195 份文档的固定语料库上采用两种协议 [10]。成员使用搜索、打开和查找工具，遵循 Li 等人提出的开放式深度研究设置 [34], , 并基于 BM25 索引进行检索；该索引在全部 830 个问题上的检索指标与官方 Lucene 索引相差不超过 0.2 分。表12使用前 100 个问题，并采用通义千问 Qwen3.6-35B-A3B，按照开放式深度研究设置中的评分提示词进行评分。

SWE-bench Verified。 我们使用前 100 个实例 [27], , 每个实例都在一个无根沙箱中运行，该沙箱使用其注册表摘要固定的官方评估镜像。成员拥有一个 shell 工具；成员停止时提交的答案就是代码库差异补丁。每个不同的补丁都会应用到一个全新的沙箱中，并由官方测试框架根据 FAIL_TO_PASS 和 PASS_TO_PASS 测试进行评分。

LongBench v2。 我们使用全部 503 个问题 [3]。上下文被拆分为多个不超过 6,000 个字符的部分，供成员搜索和读取。另会提供一份不超过 65,536 个令牌的完整文档，每个成员都会先在不使用工具的情况下对其进行讨论。单一模型参照采用基准测试的零样本模板，并将内容中间截断至 120,000 个令牌。

GPQA-Diamond。 我们使用作者发布版本中的全部 198 个问题 [47]。成员首先按照任务自身的说明，在不使用工具的情况下进行讨论；随后可以读取黑板，最后提交答案并注明置信度。单一模型参照使用模型卡中的思考设置，为每个问题采样八个答案。

MMLongBench-Doc。 从按随机种子打乱的、涵盖135份PDF文档的1,082个问题中选取的前300个问题 [39]。成员逐页工作：搜索各页面的文本层，读取页面文本并查看页面图像，每位成员最多查看40张图像。答案简短，也可以说明问题无法作答。评分时，基准测试先使用本地评判模型而非GPT-4o进行答案提取，再由基准测试的规则评分器评分。页面级协议不同于基准测试的整篇文档协议，因此这些分数与其排行榜分开列示。

FATE-M。150个问题中的前100个 [26], 每个问题都是一个需要在 Lean 4 中证明的代数定理，基于基准测试固定版本的 Mathlib。成员会搜索 Mathlib 声明，并使用检查器工具编译完整的 Lean 文件；每位成员最多编译十次，并提交一个完整的 Lean 文件作为答案。群体和集成同样

只要检查器接受任一成员的证明，问题就算解决，因此 Vote@B 与 Pass@B 相同。已发布的通过率是在全部150个问题上各取64个样本得出的，与这些单元分开。

基线。 所有组织方式共享测试框架、工具、任务提示、行动限制和评分器。并行 ReAct 运行N 彼此不沟通的成员。多智能体辩论会在每一轮向所有其他成员广播每位成员的中间工作成果。DeLM 为所有成员提供一个共享上下文，其中包含任务分解和子任务队列；为使其设计适配异步、使用工具的运行框架而做出的选择，沿用了其首个公开版本。AOrchestra 允许协调器创建并派发子智能体，并在协调器停止时终止；在代码修复任务中，其答案是最新子智能体工作区的差异补丁。解读表 12 时，有三项调整值得注意。辩论没有固定轮次数：成员在每一步都能看到其他成员的工作成果，智能体标签保持不变，并且可以在阅读这些成果之前提交；而阶梯中的辩论会将每次提交搁置一至两轮，以便成员阅读。DeLM 成员提交各自的答案，而非经过一个最终综合步骤。AOrchestra 使用同一个模型运行所有子智能体并生成其轨迹摘要，不进行模型选择。并行 ReAct 和多智能体辩论通过强制的最终步骤提交答案，而 DeLM 和 AOrchestra 则提交截至当时记录的答案，因此未记录答案的成员不会对投票作出贡献。所有组织方式均按照相同的承诺规则评分；当一次运行没有记录明确置信度时，该规则就简化为普通投票。在 SWE-bench Verified 上，没有任何规则会汇集补丁，因此承诺答案是某一位成员的补丁，Vote@B 与一名独立成员的准确率相差不超过 0.01。

K. 补充结果

K.1. 阶梯的支持性结果

第 5.2 节的比较涵盖了六个基准测试子集和审计任务中的 78 个通义千问Qwen3.5-9B 群体单元。在两个参照协议不同的五个子集中，没有任何单元在完成令牌数、总令牌数、成员数或调用次数方面超过 Pass@B，而表 12 逐一列出了先前方法的单元。在审计任务中，未训练的 16 人 9B 群体里，并行 ReAct 每次运行报告的不同配对多于黑板；而单个黑板成员召回的预设矛盾多于 ReAct 成员。黑板的成员彼此协作，并趋于认同最先发布的内容。Shapley 贡献（第 4.3 节）消除了抄袭的动机，因为重复发现会由报告者共同分摊。

其他家族在 56 项任务子集上的表现。 在设有规划窗口的情况下，在另一个批次的 56 项 HiddenBench 任务中（每项任务恰有四条私有事实），GLM-4.5-Air 成员几乎从不转移预算，而 gpt-oss-120b 成员在超过三分之一的记录中转移预算，但很少形成枢纽。允许转移并未显著改变任何一个家族的准确率。由于协议、任务集和批次均与表3b不同，因此不对这两组结果进行比较。

K.2. Shapley 贡献详解

训练设置。 信用研究在矛盾审计上训练 Qwen3.5-4B，每个步骤处理八个问题，以群体并集的精确配对 F1 值作为群体得分 s；每个成员都可以查看语料库中的所有片段。图 13 展示了置信度门控得分 $c = 0.25$，表 13 的最后一行则使用 $c = 0.5$；两者均用于比较，与表 5 的得分不同。成员以四人一组进行训练，并以四人组和八人组进行评估。每个问题进行八次 rollout，每个成员最多采取 60 个动作和步骤；每次调用最多生成 8,192 个令牌，训练序列长度不超过 32,768 个令牌。目标函数是裁剪策略梯度，包含 $\epsilon_{\text{low}} = 0.2$ 和 $\epsilon_{\text{high}} = 0.28$，不含 KL 项；每个 rollout 批次进行四次小批量更新，学习率为 10^{-6}，梯度裁剪阈值为 1.0；重要性权重截断至 2，以校正与前一个优化器步骤重叠的 rollout。Shapley 贡献使用 $\lambda = 0.25$和 $\gamma = 0.5$。训练在四块 H200 GPU 上进行，其中两块用于训练，两块用于推理；每项比较均在 16 个留出语料库和两个随机种子上配对进行，并基于 4,000 次重采样计算 95% 配对自助法区间。

单次运行中的信用分配规则。图13展示了单次运行中的三种信用分配规则。对于共享发现，自身成功规则会向两位报告者全额支付奖励。团队奖励规则给只提出错误主张的成员的奖励，与发现矛盾的成员一样多。Shapley 贡献会拆分共享发现的奖励，并让每项错误主张的作者承担相应代价。在表 5 所列语料库和随机种子上单独评估时，按照自身成功规则训练的智能体到第 40 步时，F1 值提高了一倍以上，而其所在群体的表现则低于未训练组。

按置信度筛选主张。 使错误主张代价低廉的奖励会促使系统转向提交更多但精确率更低的主张，正如命题 2 所预测的那样（表13、图 14b）。较大的惩罚会逆转这种趋势，同时也会抑制正确发现，

表12：与等预算投票对比的既有多智能体方法。使用通义千问3.5-9B成员时，多智能体辩论的承诺答案准确率减去 Vote@B 的差值（按完成令牌计） [14], DeLM [40] 和 AOrchestra [48]，（后者自行设定群体规模），以及由审阅者从成员答案中进行选择的黑板。 $^{\downarrow}$ 和 $^{\uparrow}$：95%区间低于或高于零； †：在该子集的部分数据上进行评估。

多智能体辩论

基准测试	N	Δ投票
GPQA-Diamond	4	-0.005 ± 0.013
GPQA-Diamond	8	$+0.008 \pm 0.015$
GPQA-Diamond	16	$+0.022 \pm 0.016$
BrowseComp-Plus	4	$-0.055^{\dagger} \pm 0.036$
BrowseComp-Plus	8	$-0.178^{\downarrow} \pm 0.032$
BrowseComp-Plus	16	$-0.163^{\downarrow} \pm 0.031$
LongBench v2	8	-0.016 ± 0.014
MMLongBench-Doc	16	-0.016 ± 0.017
SWE-bench Verified	8	$-0.173^{\downarrow} \pm 0.025$
FATE-M	4	-0.057 ± 0.029
FATE-M	8	-0.045 ± 0.028

DeLM

基准测试	N	Δ投票
GPQA-Diamond	4	$+0.022 \pm 0.014$
GPQA-Diamond	8	$+0.015 \pm 0.011$
GPQA-Diamond	16	$+0.022 \pm 0.015$
BrowseComp-Plus	4	$-0.002^{\dagger} \pm 0.044$
BrowseComp-Plus	16	$-0.243^{\downarrow} \pm 0.044$
MMLongBench-Doc	16	$-0.066^{\downarrow} \pm 0.022$
SWE-bench Verified	8	$-0.506^{\downarrow} \pm 0.034$
FATE-M	4	$-0.235^{\downarrow} \pm 0.035$
FATE-M	8	$-0.304^{\downarrow} \pm 0.041$

AOrchestra

基准测试	Δ投票
GPQA-Diamond	$+0.021 \pm 0.020$
BrowseComp-Plus	$-0.075^{\dagger} \pm 0.046$
MMLongBench-Doc	$-0.112^{\downarrow} \pm 0.022$
SWE-bench Verified	$-0.360^{\downarrow} \pm 0.043$

带审阅器的黑板（我们的方案）

基准测试	N	BB	+ 9B 审阅者	+ 27B 审阅者
GPQA-Diamond	16	$+0.013 \pm 0.014$	$+0.010 \pm 0.018$	$+0.050^{\uparrow} \pm 0.019$
BrowseComp-Plus	16	$+0.016 \pm 0.026$	$+0.113^{\uparrow} \pm 0.036$	$+0.131^{\uparrow} \pm 0.038$
LongBench v2	8	-0.013 ± 0.012	$+0.016 \pm 0.015$	$+0.037^{\uparrow} \pm 0.015$
MMLongBench-Doc	16	$+0.013 \pm 0.017$	-0.025 ± 0.019	$+0.044^{\uparrow} \pm 0.018$

成员	其提交中的声明	自身成功 $s_x(o_i)$	团队奖励 R	Shapley 贡献 ϕ_i
1	p_1 ✓ p_2 ✓	$+0.500$	$+0.250$	$+0.375$
2	p_1 ✓ w_1 ✗	$+0.188$	$+0.250$	$+0.063$
3	w_2 ✗ w_3 ✗ w_4 ✗	-0.188	$+0.250$	-0.188
信用总和（群体结果 $R = 0.250$）		$+0.500$	$+0.750$	$+0.250$

图 13：三条规则下单次运行的信用，以置信度门控得分为例（四项真实发现， $c = 0.25$，群体结果为 $R = 0.25$）；表 5 在精确配对 F1 上进行训练。自身成功会将同一发现 p_1 的全部奖励支付给成员 1 和成员 2；团队奖励会给只提出错误主张的成员 3 支付与其他成员同等的奖励；Shapley 贡献分配 p_1，将每条错误主张计入其作者名下，总和为 R。

表 13：各项奖励对应的付出。审计任务上的 4B 训练 rollout：每次运行的主张数量及其在运行早期和后期的精确率，依据训练 rollout 而非留出语料进行测量。 q^*：再提出一条主张可提高期望奖励所需达到的置信度。 †第 0 步和第 40 步的留出精确率， $N = 4$。

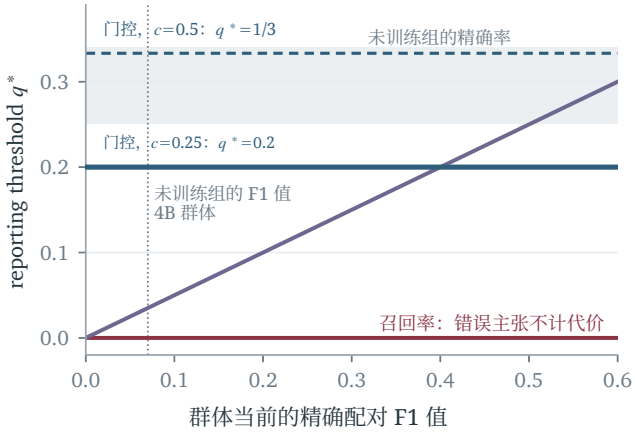
			步数		每次运行的主张数		精确率	
奖励	信用	q^*	早期	late	早期	late	早期	late
片段覆盖率	共享 + 流量	0	1	41	2.1	17.7	0.33	$0.12^{\dagger}$
精确对召回率	自身成功	0	1–5	27–31	0.8	36.0	0.34	0.05
精确对召回率	共享	0	1–3	8–10	3.0	4.1	0.30	0.24
精确配对 F1	共享 +流量	F1/2	1–5	6–10	3.6	4.8	0.25	0.20
门控主张， $c = 0.5$	共享 + 流量	1/3	1–5	10–12	3.1	0.8–1.8	0.28	0.42–0.54

表 14：沙普利份额与完整信用。分布式设置下的 Qwen3.5-4B， $N = 4$；相对于未训练模型的留出差异（单位为百分点，附 95%自助法区间）。**粗体**：区间不包含零。

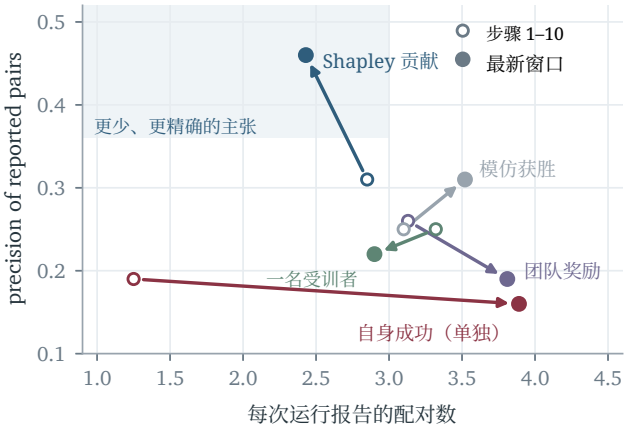
方法	Step	Δ召回率	95%置信区间	ΔF1	95%置信区间
Shapley 贡献（份额 +流量）	20	**+1.9**	[+0.8, +3.1]	**+3.0**	[+1.2, +5.2]
不含流量的Shapley归因（仅份额）	20	**+2.2**	[+0.5, +4.1]	**+2.5**	[+0.2, +5.0]
Shapley 贡献（共享 +流量）	40	**+3.8**	[+1.2, +6.4]	**+4.6**	[+1.1, +8.4]

因为 4B 群体很少能达到三分之一的置信度阈值。

份额与流量。 在分布式设置中，每个成员只能打开代码片段中属于自己的份额，仅使用沙普利份额即可重现完整信用的增益（表 14），因此 4B 处测得的收益来自这些份额本身。因此，它与未训练组之间的差异不同于表 5 中的差异，后者使用的是共享语料库。



(a) 报告阈值（命题 2）。



(b) 五种贡献分配规则下的训练动态。

图 14：贡献分配规则和置信度阈值为何重要。（a）再增加一项主张便能提高期望奖励时，该主张为正确的最低概率（命题2）。**(b)** 4B 训练期间每次运行的主张数量及其精确率（空心标记：前十个步骤；实心标记：最近窗口）。

L. 在 PlanBench、HiddenBench 和 Hanabi 上进行递归训练

共同设计。 每次运行都从四个由 vLLM 提供服务的 Qwen3.5-9B 智能体开始 [29]。每一代都从其前代的全精度权重继续，并主要使用其前代最新的已接受输出进行训练。

PlanBench。四个智能体各通过四轮调用编写积木世界计划：草案、自我检查、读取其他草案后进行修订，以及对其他计划进行匿名评审。群体提交获得最高正面评价占比的最终计划。训练、选择和确认问题均由我们自己的生成器生成，该生成器会排除所有官方 PlanBench 问题及其仅通过重命名积木得到的问题；计划模拟器会接受每个计划。问题大多涉及五个积木。每次调用可使用 8,192 个令牌，供思考和作答共同使用。选择集包含 300 个问题，每个确认切片包含 800 个问题。当 $d \geq 1.645$标准误 和 $d > 0$满足条件时，候选项即会发布；训练在发布四代后停止。

HiddenBench。 四个智能体进行三轮辩论，并根据各自声明的赌注加权后提交答案。训练和选择使用隐藏档案任务，这些任务由基础模型按照 HiddenBench 的格式写入 [33]；当某一轮的承诺答案与至少三名基础模型成员在每个事实上的答案一致，且通过协议的检查时，该轮才会被接受。选择集包含 200 个隐藏档案任务和 100 个对照任务，使用两个随机种子运行；每轮还会抽取新的确认种子。当 d超过 0.026，且其在对照任务上的准确率保持在以基础模型为基准设定的下限之上时，候选项会被发布。T3 在相同上限下增加了一次匹配的单智能体自训练运行。T3k 再次从基础模型开始，进行了四项改变：将对照任务保留在训练行中，使用半步骤和完整步骤的候选项，将决策令牌仅计数一次，并按照调用结构对答案值加权。它没有对应的自训练版本。两次运行均在八轮后停止。

Hanabi。 四个智能体使用固定的行动词汇表进行合作式纸牌游戏 [4]，每次行动进行一次调用。对于每一局训练牌局，群体进行四场游戏，并将得分高于该牌局平均值的最佳一场保留为精英对局。每轮抽取 300 个全新的选择牌局和 600 个全新的确认牌局；当 d –标准误 > 0 达到挑战得分上的要求时，候选项会被发布。非递归对照运行在首次发布后分支，并使用固定代的游戏进行训练。训练在发布六代后停止。表15列出了三次运行中每轮的发布决策。

披露。 T3、T3k 和 Hanabi 是在读取其他基准测试的早期运行结果后设计的；图 11 所引用的 HiddenBench 任务也已事先读取。T3k 的设计，以及 PlanBench 的领域和令牌上限，都是在对同一系列进行早期测量后选定的。PlanBench 缺少单智能体自训练运行，因此其结果将集体学习与个体学习合并呈现。在 PlanBench 上，截断调用所占比例随代数递减；每一代的令牌数均与其准确率一并报告。

表 15：至少有四次发布的三次运行的发布决策。每轮都会在新的确认切片上，按问题配对比较选定候选项与当前最佳项；d 是配对差值及其标准误。PlanBench 在$d \geq 1.645$标准误时发布；HiddenBench (T3k) 在 d 超过 0.026 时发布；Hanabi 在 $d -$ 标准误 > 0时发布。PlanBench 和 HiddenBench 以群体解决的问题比例计分，Hanabi 的挑战得分满分为 25。运行日志中未记录的值标记为（–）。

	PlanBench		HiddenBench (T3k)		Hanabi	
轮次	d (标准误)	决策	d (标准误)	决策	d (标准误)	决策
1	$+0.295$ (0.020)	G1	$+0.203$ (–)	G1	$+0.678$ (0.045)	G1
2	$+0.080$ (0.014)	G2	$+0.044$ (0.026)	G2	$+0.04$ (0.051)	flat
3	$+0.024$ (0.012)	G3	$+0.001$ (0.022)	flat	-0.085 (0.054)	flat
4	$+0.021$ (0.010)	G4	$+0.044$ (–)	G3	-0.012 (0.052)	flat
5			$+0.065$ (0.020)	G4	–	flat
6			$+0.001$ (0.021)	flat	$+2.363$ (0.075)	G2
7			$+0.046$ (0.021)	G5	$+1.102$ (0.087)	G3
8			$+0.015$ (0.016)	flat	$+0.208$ (0.077)	G4
9					$+0.087$ (0.074)	G5
10					$+0.125$ (0.075)	G6

M. 计算与可复现性

阶梯实验的推理在一个集群的 H200 GPU 上，以及两个各配有八块 GPU 的节点的 A800 GPU 上运行；vLLM 服务器采用一或二路张量并行。由四个 27B 成员组成、包含 448 条记录的单元，在一个四 GPU 群组上需约 9 至 26 小时，具体取决于协议；由十个单元和样本池组成的 27B GPQA-Diamond 批次，在八块 H200 GPU 上耗时七个半小时。对一代 27B 模型进行监督微调，在八块 A800 GPU 上约需 75 分钟；对一代 9B 模型进行监督微调，在四块 GPU 上约需 17 至 27 分钟。Shapley 贡献度运行使用 H200 GPU：一个 4B 训练步骤在四块 GPU 上需 10 至 14 分钟。每次运行均基于代码的冻结快照执行，并记录相应的提交。基准测试数据和模型权重均固定到特定修订版本，并在使用前通过内容哈希验证；除表格另有说明外，每个样本池都在其所评分单元对应的平台和批次中抽取：例外是测试时自写提示词单元的描述性跨批次读数。