TTS(测试时扩展)对多智能体协作的提升

Two Heads are Better Than One: Test-time Scaling of Multi-agent Collaborative Reasoning

Can Jin
Rutgers University
can.jin@rutgers.edu

Hongwu Peng
University of Connecticut
hongwu.peng@uconn.edu

Qixin Zhang
Nanyang Technological University
qixinzhang1106@gmail.com

Yujin Tang
Sakana AI
yujintang@sakana.ai

Tong Che‡
NVIDIA Research
tongc@nvidia.com

Dimitris N. Metaxas‡
Rutgers University
dnm@cs.rutgers.edu

李宏扬
2025.12.17

研究动机

- 人工智能的目标： 创造能在现实世界无缝运行的智能自主体。
- LLM 智能体的局限： 单智能体 LLM 难以应对复杂任务的内在复杂性 。
- MAS 的潜力： 通过协作解决数学、软件开发、科学发现等任务 。
- 核心问题（本文着力解决）： 如何有效扩展（Scaling）MAS 中的协作和推理能力（即 TTS for MAS）？

背景知识

- LLM 智能体
- 最新的工作已经能将 LLM 的能力扩展到独立推理和理解之外，使它们能够作为多智能体运行，与其他环境、工具和智能体进行交互以执行复杂任务 。这些多智能体系统（MAS）集成了各种技术 ，包括 CoT 提示 、迭代细化 、自我改进 和外部工具使用 ，以支持多步骤决策和长远期规划 。它们已成功应用于数学推理 、软件工程 和科学发现等领域 。智能体框架通常使用少样本提示 和指导性推理 等技术来组织与 LLM 的交互 ，依赖于模型的上下文学习能力 。

背景知识

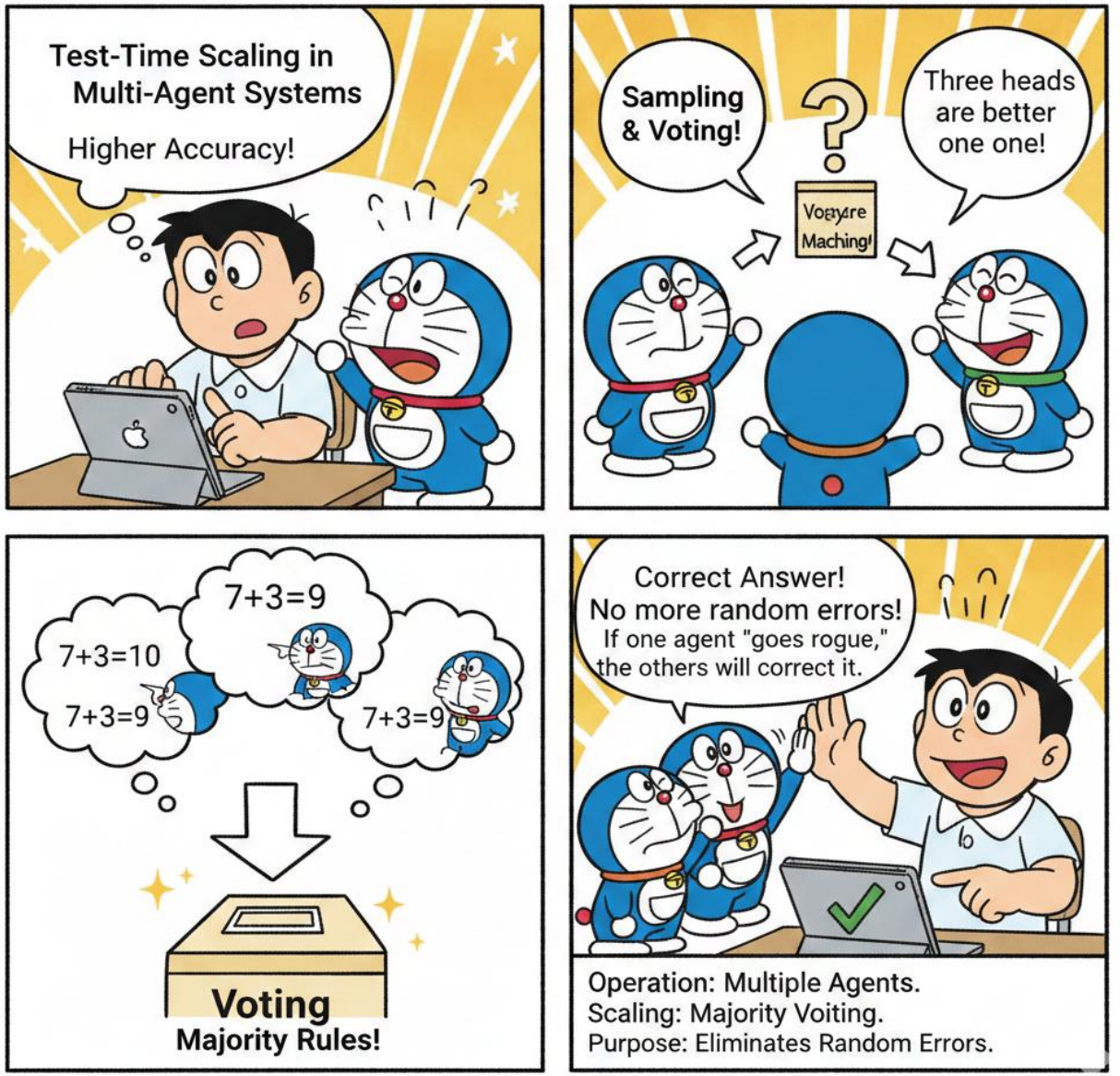
- 多智能体系统（Multi-agent systems, MAS）
- 多智能体系统（Multi-agent systems, MAS）是在人工智能（AI）领域中，用于解决复杂现实世界任务的一种系统，基于大型语言模型（LLMs）构建的系统。
- 作用： 它们提供了一条道路，用于解决单智能体系统往往难以应对的复杂现实世界任务。
- 机制： MAS 通过利用多个 LLM 智能体之间的协作交互来有效应对各种任务。
- MAS 框架通常会涉及多种角色进行协作，例如：
  - 专家招聘者（Expert Recruiter）：负责识别和分配合适的专家。
  - 问题解决者（Problem Solver）：被招募的专家，通过多轮讨论提出并迭代细化解决方案。
  - 执行者（Executor）：运行必要的代码或调用外部工具。
  - 评估者（Evaluator）：审查解决方案和结果，提供反馈。

背景知识

- 测试时扩展（Test-Time Scalling）
- 在推理阶段增加计算量，通过**让多个智能体协作、辩论或搜索，来解决单体模型无法解决的复杂问题**
- 已开发出多种方法，通过利用测试时扩展（TTS）来改进 LLM 中的推理 。最近的工作探索了包括分层假设搜索等技术 ，它通过结构化探索实现归纳推理 ，以及推理时工具增强 ，它通过允许模型与外部环境交互来增强下游性能 。其他方法侧重于内部机制 ，例如以无监督方式学习思维标记（thought tokens） ，使模型能更好地利用扩展的推理序列 。

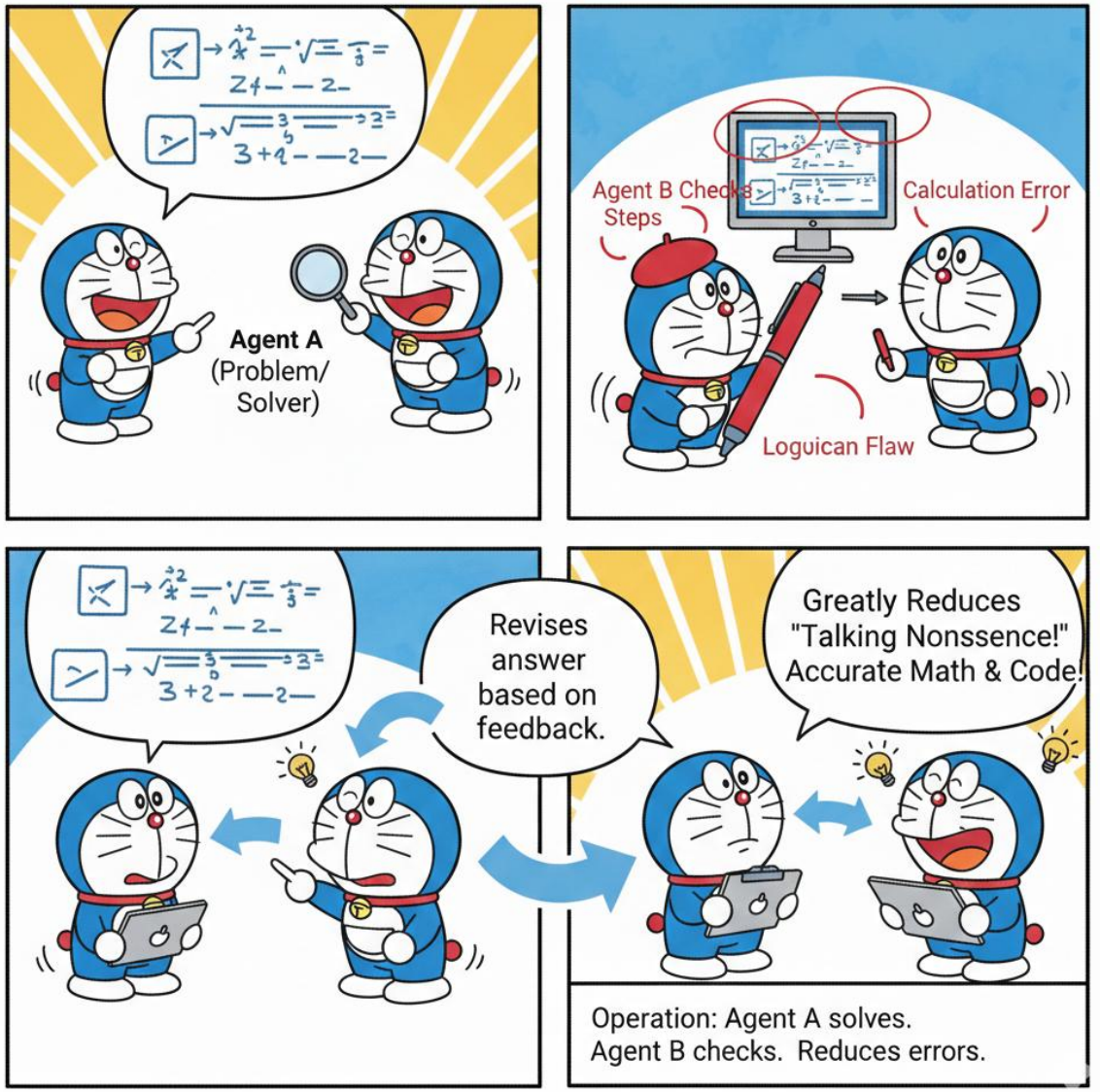
背景知识

- 测试时扩展（Test-Time Scalling）
- 三种主要的“扩展”模式
- 在多智能体系统中，测试时扩展通常通过以下三种方式“换取”更高的精度：
- A. 采样与投票 (Sampling & Voting) —— “三个臭皮匠”这是最基础的扩展形式（类似于视觉里的 TTA）。
  - 操作： 让多个智能体（或同一个智能体多次）独立回答同一个问题。
  - 扩展： 使用“多数投票”（Majority Voting）或“加权投票”来决定最终答案。
  - 作用： 消除单次推理的随机性错误。如果一个智能体“发疯”了，其他智能体会把它纠正回来。



背景知识

- 测试时扩展（Test-Time Scalling）
- 三种主要的“扩展”模式
- B. 验证与批判 (Verification & Critique) —— “红蓝对抗”这是多智能体最擅长的领域。
  - 操作：Agent A（做题家）： 生成一个解题步骤。Agent B（判卷老师）： 检查 A 的步骤，指出哪里有逻辑漏洞或计算错误。Agent A： 根据 B 的反馈修改答案。
  - 扩展： 这个循环可以进行 N 轮。
  - 作用： 极大地减少了“一本正经胡说八道”的现象，特别是在数学、代码生成领域。

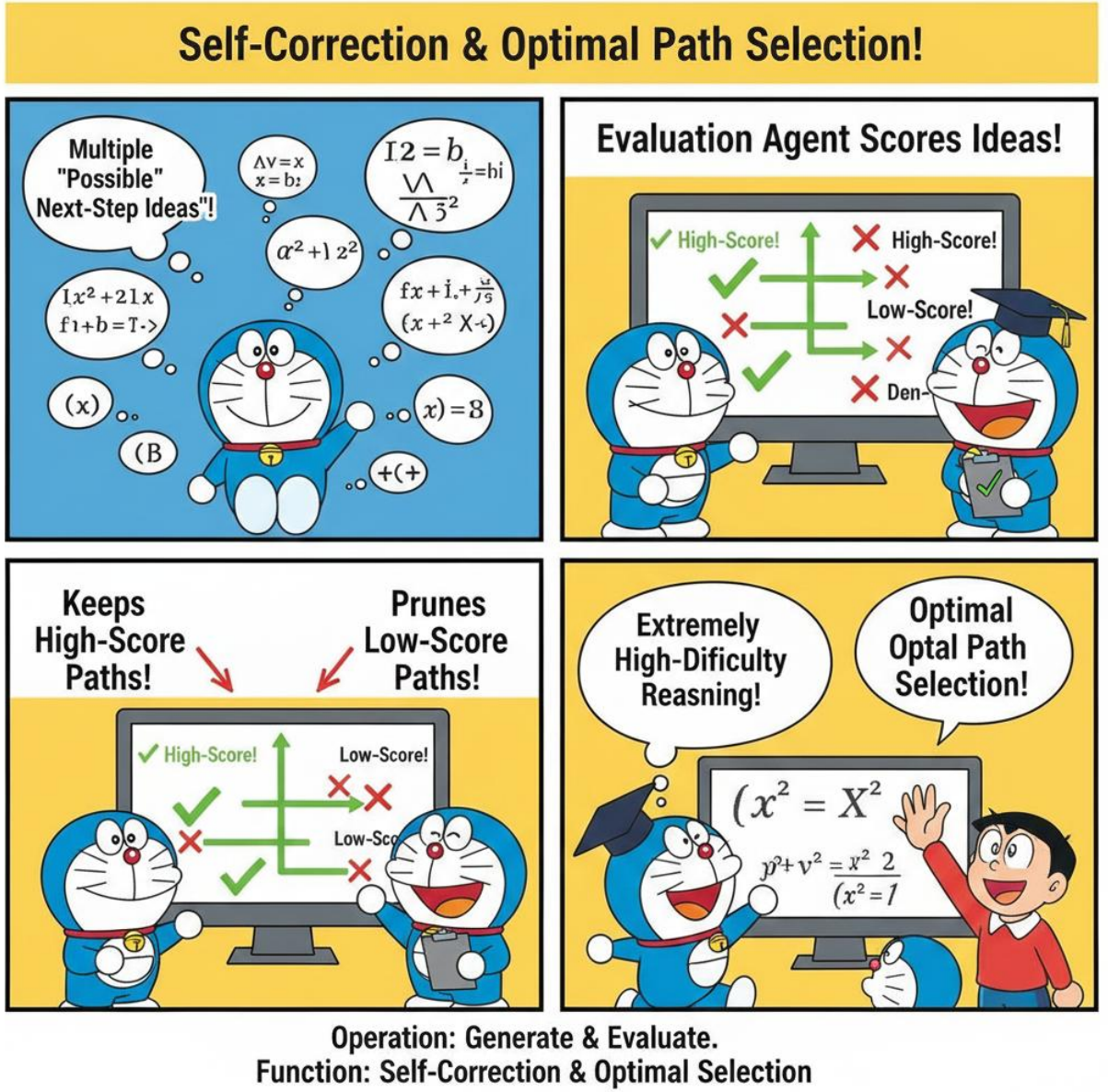


背景知识

- 测试时扩展（Test-Time Scalling）
- 三种主要的“扩展”模式
- C. 树搜索与规划 (Tree Search / Process Reward) —— “思维树”--目前最前沿的方向

（类似 OpenAI o1 背后的逻辑）

  - 操作： 智能体不直接生成最终答案，而是生成多个“下一步的可能想法”。
  - 扩展： 另一个“评估智能体”会对这些想法打分，系统会保留高分的路径，剪掉低分的路径（类似于下围棋时的 AlphaGo）。
  - 作用： 能解决极高难度的推理问题，因为智能体在每一步都在进行“自我纠错”和“最优路径选择”。



背景知识

- 测试时扩展（Test-Time Scalling）

- 在最受研究的扩展范式中，包括并行和顺序 TTS 方法 。并行方法独立生成多个解决方案候选，并使用评分标准（如多数投票或基于结果的奖励模型）选择最佳方案 。相反，顺序方法将每次新的尝试都以前一次为条件 ，允许基于先前输出来进行迭代细化 。连接这些策略的是基于树的技术，例如蒙特卡洛树搜索（MCTS）和引导式集束搜索（guided beam search），它们通过分支和评估实现结构化探索 。许多这些方法的核心是奖励模型 ，它为生成提供反馈信号 。这些可分为结果奖励模型 ，用于评估完整的解决方案 ，或过程奖励模型 ，用于评估中间推理步骤 ，指导模型走向更有效的推理路径 。

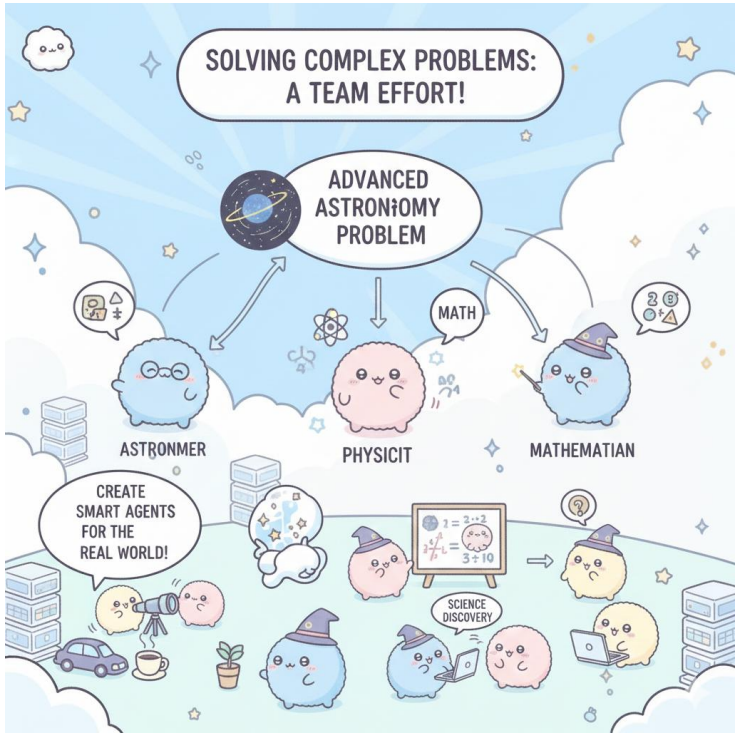
2.1多智能体协作推理数据的自动生成

2.1.1基于难度、多样性和跨学科性进行问题抽样

使用 Simple-Scaling 的完整数据集，包含多个来源的各种问题 。这些问题涵盖了物理学、几何学、数论、生物学和天文学等多个领域。

使用 DeepSeek-R1 来确定解决每个问题是否需要跨学科知识 ， 并排除了 DeepSeek-R1 使用少于 1024 个标记回答的问题 。

例如：解决一个高级天文学问题可能需要天文学家、物理学家和数学家的知识输入 。



2.1多智能体协作推理数据的自动生成

● 2.1.2 生成多智能体协作推理轨迹（trace）

- 为了生成协作推理轨迹，文章采用开源 MAS 框架和推理模型，主要是 AgentVerse 和 DeepSeek-R1 ，来处理先前选择的问题 。该 MAS 框架涉及多个角色 ： 专家招聘者（Expert Recruiter）、问题解决者（Problem Solver，例如科学家和工程师）、执行者（Executor，例如质量保证工程师）和评估者（Evaluator，例如产品经理）。它们按以下方式协作：

1.专家招聘者识别并分配合适的专家来执行任务（分配的智能体总数是固定的）

2.专家充当问题解决者，通过与预定的批评迭代次数进行多轮讨论 ， 提出并迭代细化解决方案，直到到达预设的最大迭代次数，然后将得到的方案传递给执行者

3.执行者运行所需的代码或在需要时调用外部工具

4.评估者审查解决方案和执行者产生的结果 ， 为可能的细化提供反馈 。

这四步会反复迭代执行直到到达最大迭代次数

在这个过程中，每个智能体的输入——输出会被收集，从而生成推理轨迹

推理轨迹样例

Multi-agent Collaborative Reasoning Trace

Expert Recruiter

Prompt: You are leading a group of experts to solve: ... Find the greatest integer that does not exceed $\frac{x^2}{100}$. Recruit two experts with different specialties to best solve the problem...
Response: <think> ... I need to identify experts best suited for this geometry task... </think>1. A geometric analysis expert. 2. An algebraic modeling expert.

Problem Solver 1 (Geometric Expert)

Prompt: ... Solve the following accurately: One base of a trapezoid is 100 units longer than the other... Find the greatest integer that does not exceed $x^2/100$...
Response: <think> ... Let's understand the structure: the segment splits the area in a 2:3 ratio. We need the length x such that both halves have equal area... </think>...Using RMS to find x, we get:

$$\frac{x^2}{100} = \frac{18125}{100} = 181.25 \Rightarrow \boxed{181}$$

Problem Solver 2 (Algebraic Expert)

Prompt: ... Given the prior solution, critique its correctness. Do not solve again. If you agree with the final answer, append `[Agree]`...
Response: <think> ... Reviewing the solution: the answer is correct. However, explanation for using RMS could be clearer... </think>... The reasoning is sound and the steps are accurate. `[Agree]`

Evaluator

Prompt: You are an experienced science teacher. Evaluate the final solution from both solvers for correctness...
Response: <think> ... Let me verify each step. Yes, the use of RMS and the calculation are both appropriate... </think>... **Correctness:** 1 ...

2.2通过监督微调增强 LLM 协作推理

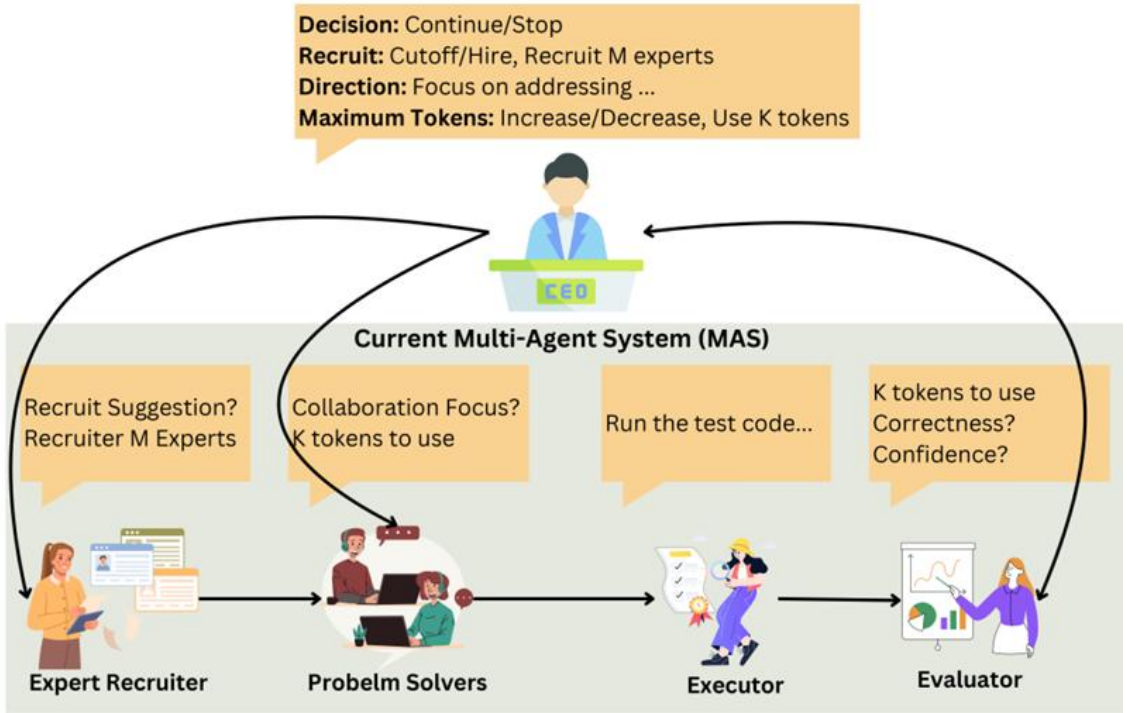
- 受到 Simple-Scaling 的启发【该研究表明 LLM 的长 CoT （Chain-of-Thought Prompting）推理能力可以通过对详细推理轨迹进行 SFT (Supervised Fine-Tuning)来发展】
- 用 M500 数据集对 LLM应用SFT，有监督微调的目的是最小化损失函数

$$\mathcal{L}_{\text{SFT}} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\in \text{M500}} \left[-\frac{1}{|\mathbf{y}|} \sum_{t=1}^{|\mathbf{y}|} \log P_f(\mathbf{y}_t \mid \mathbf{x}, \mathbf{y}_{<t}) \right]$$

- 其中 $P_f(y_t \mid x, y_{<t})$ 表示模型 f 在给定输入 x 和先前标记$y_{<t}$ 的情况下赋予标记 y_t 的概率 。
- 我们确保 同一问题 Q 的所有推理轨迹 $\{(x_i, y_i)\}_{i=1}^{n}$ 在同一批次中分组，并按照 MAS 中的原始生成顺序排序 。

2.3 自适应测试时扩展

- 提出了一种针对 MAS 的自适应 TTS 策略，通过引入一个专用的 “CEO”智能体来实现，该智能体根据给定任务的持续进展动态管理协作和资源分配
- CEO 智能体评估问题、当前解决方案状态、评估反馈和可用资源，以确定是否应接受提议的解决方案或需要进一步细化



与具有固定智能体数量、迭代限制和推理深度的静态 MAS 配置不同之处在于，该自适应方法允许 MAS 动态调整其设置 。

对实验方法的评估

- 为了进行全面评估，我们重点关注三个关键领域
- 通用理解（General Understanding）
- 数学推理（Mathematical Reasoning）
- 代码编写（Coding）
- 训练和评估。 使用 M500 数据集对 Qwen2.5 进行 SFT，训练 5 个 epoch，学习率为 1e-5，得到了的模型 M1-32B 268。训练是在 LLaMA-Factory 框架 内使用 FlashAttention 和 DeepSpeed在 8 块 NVIDIA A100 GPU 上进行的。

实验结果

在 AgentVerse 框架内使用强大的推理和非推理模型，对通用理解、数学推理和代码编写任务的性能比较 。方法在所有任务上都比 Qwen2.5 和 s1.1-32B 有实质性改进 ， 并在 MATH-500 和 MBPP-S 上取得了与 O3-mini 和 DeepSeek-R1 相当的性能 ， 证明了其在增强 MAS 中协作推理方面的有效性

CEO 智能体的引入持续改善了 M1-32B 在所有任务上的性能

Model	General Understanding		Mathematical Reasoning		Coding	
	GPQA	Commongen	AIME2024	MATH-500	HumanEval	MBPP-S
Non-Reasoning Models						
Qwen2.5	50.2	96.7	21.1	84.4	89.0	80.2
DeepSeek-V3	**58.6**	**98.6**	**33.3**	**88.6**	89.6	83.9
GPT-4o	49.2	97.8	7.8	81.3	**90.9**	**85.4**
Reasoning Models						
s1.1-32B	58.3	94.1	53.3	90.6	82.3	77.4
DeepSeek-R1	**75.5**	97.2	78.9	**96.2**	**98.2**	91.7
o3-mini	71.3	**99.1**	**84.4**	95.3	97.0	**93.6**
M1-32B (Ours)	61.1	96.9	60.0	95.1	92.8	89.1
M1-32B w. CEO (Ours)	62.1	97.4	62.2	95.8	93.9	90.5

MAS 中的“顿悟时刻”（Aha Moment）

- 在使用 M1-32B 的 MAS 中，我们观察到智能体有时会表现出涌现行为，主动帮助验证和改进协作过程，即使这不是明确要求的 。

Aha Moment in MAS

Question

Let ABC be a triangle inscribed in circle ω. The tangents to ω at B and C intersect at point D, and $\overline{AD}$ intersects ω again at P. If $AB = 5$, $BC = 9$, and $AC = 10$, AP can be written as $\frac{m}{n}$ in lowest terms. Find $m + n$.

Problem Solver

... Place $B = (0, 0)$, $C = (9, 0)$, compute power of D: $DB^2 = \left(\frac{225}{22}\right)^2$. Use the power of a point: $DA \cdot DP = DB^2$, solve for DP, then compute $AP = DP - DA$. Final result: $AP = \frac{22}{1} \Rightarrow m + n = \boxed{23}$.

CEO

<think>...Wait, let's see what the solution did. They said $DA = \frac{325}{22}$, $DB^2 = \left(\frac{225}{22}\right)^2$. Then:

$$DP = \frac{(225/22)^2}{325/22}, \quad AP = DP - DA = \frac{2025}{286} - \frac{325}{22} = \frac{-1100}{143}$$

But length can't be negative, which indicates a mistake in the sign. This suggests that the assumption about the position of P is wrong. Maybe P is between D and A. So $AP = \frac{1100}{143} = \frac{100}{13} \approx 7.6923$

The solution says $AP = 22$. That's impossible. The correct value is $AP = \frac{100}{13}$, hence $m + n = \boxed{113}$... </think>

Decision: **Continue**
Recruit Number: 2
Direction: 1) Verifying the power of point application 2) Re-examining the geometric configuration ...
Maximum Tokens: 8192

Problem Solver

... Power of point D calculation: $DA \cdot DP = DB^2$
Where: $DA = \frac{325}{22}$ (distance from D to A), $DB = \frac{225}{22}$ (tangent length from D to circle). Solving for DP:

$$DP = \frac{DB^2}{DA} = \frac{\left(\frac{225}{22}\right)^2}{\frac{325}{22}} = \frac{2025}{286}$$

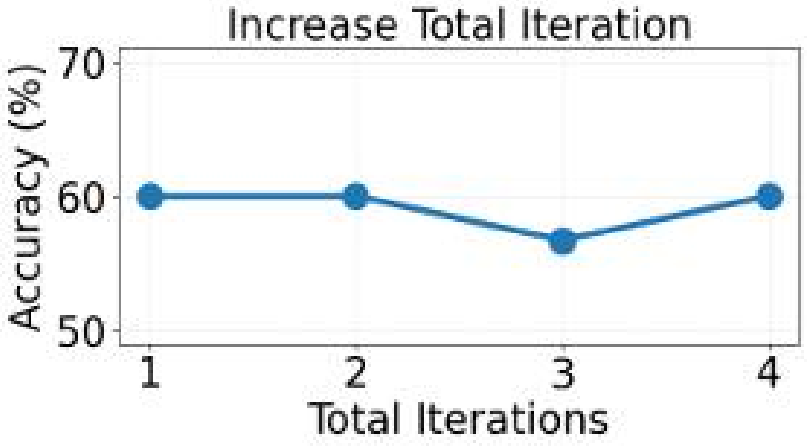
Critical geometric relationship: Point P lies between D and A, so: $AP = DA - DP = \frac{325}{22} - \frac{2025}{286} = \frac{100}{13}$

Final result: $AP = \frac{100}{13} \Rightarrow m + n = \boxed{113}$

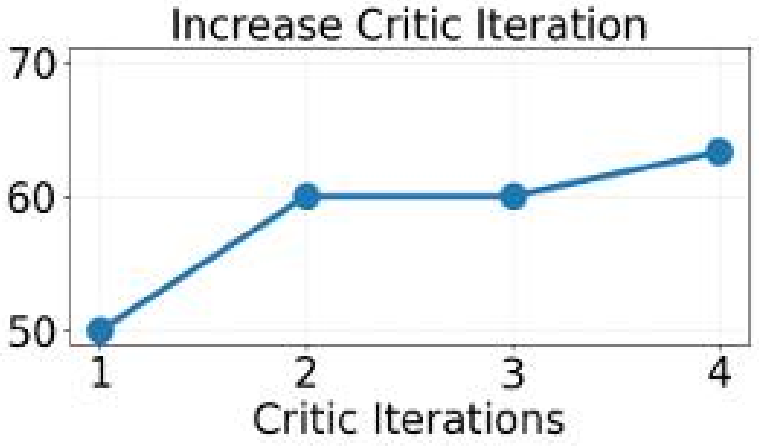
MAS 中协作和推理的扩展

- 通过系统地调整总迭代次数、批评迭代次数、总智能体数量和最大标记限制 ， 研究了协作和推理的扩展如何影响 M1-32B 在 MAS 中的性能 。

增加迭代次数



增加批评次数

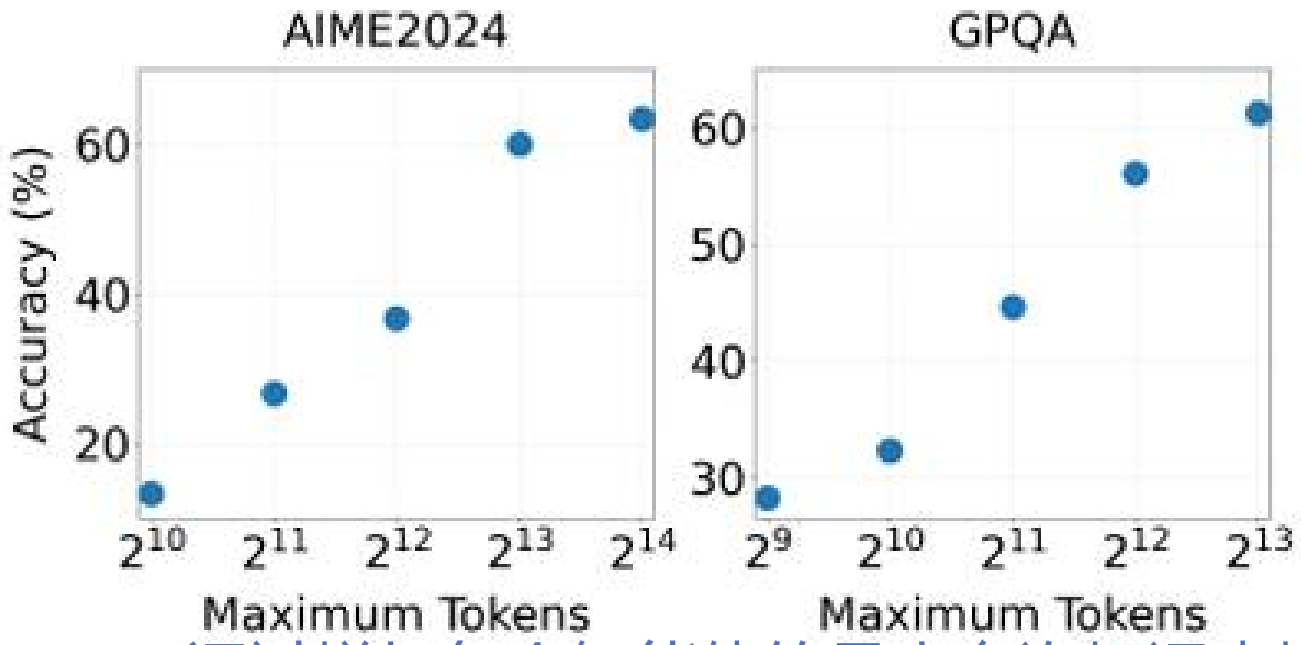


增加智能体数目

通过增加问题解决者之间的交互来增强协作，可显著提高性能 。

MAS 中协作和推理的扩展

- 通过系统地调整总迭代次数、批评迭代次数、总智能体数量和最大标记限制 ， 研究了协作和推理的扩展如何影响 M1-32B 在 MAS 中的性能 。



通过增加每个智能体的最大允许标记来增强推理能力，可以有效地提高 MAS 性能

05 结论

- 在本文中，作者引入了一种自适应 TTS 方法来增强多智能体协作推理能力 。我们通过自动生成过程构建了专门用于多智能体协作推理任务的 M500 数据集 ， 并在此数据集上微调了 Qwen2.5-32B-Instruct 模型 ， 得到了适用于 MAS 协作推理的 M1-32B 模型 。此外，我们提出了一个 CEO 智能体 ， 旨在自适应地管理协作和推理资源 ， 进一步提高了 M1-32B 在 MAS 内的性能 。广泛的实验结果表明，我们的方法显著超越了 Qwen2.5-32B-Instruct 和 s1.1-32B 模型在 MAS 中的性能。

06 我的一些想法

- 1.本文选用的“问题解决者”仅仅是未经特定领域知识，根据QA轨迹进行微调过的大模型，只是提前用prompt赋予了普通大模型“xx领域专家”的身份，让他回答问题。如果有合适的语料或专业领域知识，可以明显改进本文中对不同领域的解决效率

- 2.动态资源分配的必要性： 实验证明，最佳的推理深度（最大标记数）会因任务而异 。因此，使用像 CEO 智能体这样的自适应机制来动态调整资源分配，能够实现计算资源的最优化利用