Summary：

MAS_Arena实现了全栈式的评测基准平台，在基础的MAS和Benchmark维度外扩展了Workflow_Optimazation，Failures Attribution，Visualization等。

背景与核心问题

随着大语言模型（LLM）从单体模型向协作式多智能体系统演进，研究者面临着**碎片化的工作流**：

- **开发与评估脱节**：构建智能体（如使用 AutoGen）和在标准基准（如 GAIA）上进行评估通常需要复杂的自定义代码 。
- **黑盒评估限制**：现有基准通常只关注最终的成功率，忽略了内部的交互效率、Token 消耗和延迟等关键指标 。
- **诊断困难**：当系统失败时，难以确定是逻辑循环、工具错误还是上下文溢出导致的 。

FLEXMAS 平台架构

FLEXMAS 采用了"解耦设计"哲学，将系统分为三个核心层：

- **组合层（Composition Layer）**：
  - 提供"乐高式"的构建体验，允许用户通过组合**模型后端**（M）、**角色配置**（P）、**工具集**（T）、**存储模块**（Mem）和**交互策略**（π）来重构各种智能体架构 。
- **运行时引擎（Runtime Engine）**：
  - 作为一个**通用适配器**，支持 GAIA、AIME 和 HumanEval 等主流数据集。
  - 使用 **Docker 容器**提供安全的沙盒环境，支持代码生成和文件操作任务。
  - 内置**观察者（Observer）**模式，在不干扰逻辑的情况下捕获 API 调用和消息传递 。
- **分析层（Analysis Layer）**：
  - 提供"全栈式"诊断，包括轨迹可视化（生成节点与边的交互图）和详细的成本归因分析

主要贡献

- **模块化重构**：无需更改评估逻辑即可复现如 Camel、LLM-Debate 等多种智能体范式 。
- **白盒诊断**：通过自动化故障分类（如幻觉、最大轮次限制等）和通信拓扑图，提供深度的系统洞察 。
- **性能与效率平衡**：在衡量成功率的同时，精确统计 Token 使用量和执行成本，为实际应用部署提供参考 。

BENCH Agent

benchAgent 是一个集合了工具调用、模型调用的通用智能体，用于最简单的调用问题

重点在于构建一个高度可用，简单易用，扩展性强的智能体评测框架。

实验大纲

Method	Assigner	Math	AIME	DROP	MMLU_pro	BBH	Humaneval	gsm8k	hotpotQA	ifeval	mbpp	Gaia	accuracy	cost	time
Single															

Method	Assigner	Math	AIME	DROP	MMLU_pro	BBH	Humaneval	gsm8k	hotpotQA	ifeval	mbpp	Gaia	accuracy	cost	time
EvoAgent															
Meta-GPT															
LLM-debate															
Camel															
AutoGen															
Jarvis															
ChatEval															
MAD															