MASLab: A Unified and Comprehensive Codebase for LLM-based Multi-Agent Systems

Rui Ye[1,#] **Keduan Huang**[1,#] **Haochen Zhao**[3,#] **Xiangrui Liu**[1,#] **Yuzhu Cai**[1,#]
Tian Jin[1,#] **Jiaqi Su**[1,#] **Yue Hu**[4,#] **Zhenfei Yin**[5,6,#] **Lei Bai**[2,#,†] **Siheng Chen**[1,#,†]

[1] Shanghai Jiao Tong University [2] Shanghai AI Laboratory [3] UCLA
[4] University of Michigan [5] University of Oxford [6] The University of Sydney
[#] MASLab: https://github.com/MASWorks/MASLab

"As we teach agents to collaborate, we too should collaborate to build a unified foundation for MAS."

— MASLab Members to the MAS Community

Abstract

LLM-based multi-agent systems (MAS) have demonstrated significant potential in enhancing single LLMs to address complex and diverse tasks in practical applications. Despite considerable advancements, the field lacks a unified codebase that consolidates existing methods, resulting in redundant re-implementation efforts, unfair comparisons, and high entry barriers for researchers. To address these challenges, we introduce MASLab, a unified, comprehensive, and research-friendly codebase for LLM-based MAS. (1) MASLab integrates over 20 established methods across multiple domains, each rigorously validated by comparing step-by-step outputs with its official implementation. (2) MASLab provides a unified environment with various benchmarks for fair comparisons among methods, ensuring consistent inputs and standardized evaluation protocols. (3) MASLab implements methods within a shared streamlined structure, lowering the barriers for understanding and extension. Building on MASLab, we conduct extensive experiments covering over 10 benchmarks and 5+ models, offering researchers a clear and comprehensive view of the current landscape of MAS methods. MASLab will continue to evolve, tracking the latest developments in the field, and invites contributions from the broader open-source community.

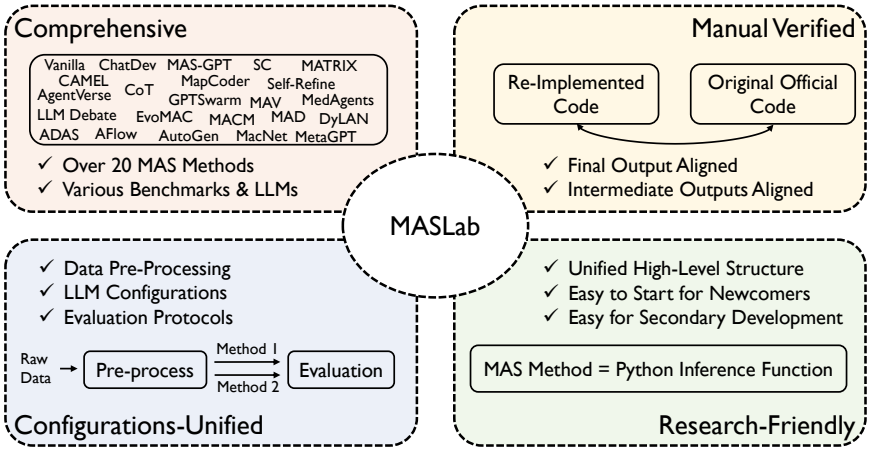

Figure 1: MASLab: A unified, comprehensive, and research-friendly codebase for LLM-based MAS.

Preprint. In Process.

1 Introduction

"The power of intelligence stems from our vast diversity, not from any single, perfect principle."

— Marvin Minsky, The Society of Mind, p. 308

Large language models (LLMs), such as ChatGPT [1], Claude [2], Deepseek-R1 [3], Qwen [4], and Llama [5], have achieved remarkable success and are being increasingly applied across a wide range of domains [6, 7, 8, 9]. However, despite their continuous advancements, a single LLM inherently faces limitations such as unreliable and random generation [10, 11], hallucinations [12, 13], and difficulty handling complex, multi-step tasks [14, 15]. These limitations hinder their ability to effectively tackle the full spectrum of real-world applications on their own.

The limitations of single LLMs have driven emerging research towards the development of LLM-based multi-agent systems (MAS) [16, 17, 18, 19], where multiple agents, each with distinct roles, contexts, and tools, collaborate to address complex tasks more effectively. This paradigm has shown great promise across a range of applications, including code generation [16, 20], mathematical problem-solving [21, 22], academic research [23, 24], and data synthesis [25, 26]. Over the past year, this field has seen rapid development, evolving from early MAS approaches that rely on manually designed, fixed systems [16, 27, 20, 28, 29, 21] to more dynamic systems where the roles and behaviors of the agents are adaptable through LLMs [18, 19, 30, 31, 32]. This ongoing evolution is steering the field towards greater automation and generalization, with the potential to create more intelligent and versatile systems.

Despite the rapid progress in LLM-based MAS, the field lacks a unified codebase that consolidates the various methods and algorithms. This gap results in several critical issues that hinder long-term advancement: **(1) Redundant effort.** Due to the absence of shared, accessible resources, researchers expend significant time reimplementing existing works, diverting effort from innovative contributions. **(2) Unfair comparison.** Varied implementation designs of individual codebases, such as differing dataset preprocessing and evaluation protocols, complicate fair and reliable comparisons across methods. **(3) High entry barriers.** Newcomers face difficulties navigating through disparate repositories, with no clear starting points. Addressing these challenges is crucial to accelerate research and promote cohesive progress in the field.

To address these challenges, we present **MASLab**, the first unified codebase for LLM-based multi-agent systems, integrating over 20 established methods (see Table 1) with standardized evaluations and a coherent high-level structure. (1) MASLab consolidates diverse research spanning multiple domains—including general tasks, coding, mathematics, and scientific research—covering representative advancements from March 2023 through March 2025. Each method integrated into MASLab has been rigorously verified by comparing step-by-step outputs with its original, official implementation, greatly reducing redundant reimplementation efforts for future researchers. (2) MASLab supports unified evaluations across a wide array of benchmarks, ensuring consistent inputs and standardized evaluation protocols. This facilitates reliable and fair comparisons, emphasizing core methodological differences rather than implementation disparities. (3) All methods are implemented within a streamlined, high-level structure, where each is encapsulated as a core inference function that processes a query and delivers the MAS response. This transparent structure explicitly highlights key methodological components, significantly lowering entry barriers and enabling researchers to easily understand, extend, and innovate upon existing approaches.

Based on MASLab, we conduct extensive experiments to compare the implemented methods, aiming to provide researchers with a clearer understanding of the current landscape of MAS approaches (see Table 2). Specifically, we evaluate over 10 benchmarks spanning diverse domains—including general question answering, mathematics, coding, science, and medicine—using various LLM backbones such as Llama, Qwen, and GPT models. In addition to accuracy, we also compare the token consumption of each method, offering a more comprehensive perspective on the trade-offs between performance and efficiency.

Table 1: Descriptions of 23 methods that MAS-Lab currently support. We show several critical perspectives of MAS methods. (1) Role: whether agents' roles in the method is fixed or dynamic. (2) Topo.: whether the topology in the method is fixed or dynamic. (3) Tool: whether the method includes tool usage. (4) Optim.: whether the method is optimizable. (5) Generalization: whether the method can generalize to handle diverse tasks.

No.	Methodology	Venue	Role	Topo.	Tool	Optim.	Generalization
Single-Agent Baselines							
①	Vanilla LLM	-	Fixed	Fixed	No	No	Yes
②	CoT [33]	NeurIPS 2022	Fixed	Fixed	No	No	Yes
Multi-Agent Systems for General Tasks							
③	CAMEL [17]	NeurIPS 2023	Fixed	Fixed	No	No	Yes
④	AutoGen [34]	ICLR-W 2024	Fixed	Fixed	Yes	No	Yes
⑤	Self-Consistency [35]	ICLR 2024	Fixed	Fixed	No	No	Yes
⑥	AgentVerse [29]	ICLR 2024	Dynamic	Fixed	No	No	Yes
⑦	LLM Debate [27]	ICML 2024	Fixed	Fixed	No	No	Role Modification
⑧	GPTSwarm [32]	ICML 2024	Fixed	Dynamic	Yes	Yes	Re-Optimization
⑨	DyLAN [31]	COLM 2024	Fixed	Dynamic	No	No	Role Modification
⑩	MAD [28]	EMNLP 2024	Fixed	Fixed	No	No	Role Modification
⑪	Self-Refine [36]	NeurIPS 2024	Fixed	Fixed	No	No	Yes
⑫	MacNet [37]	ICLR 2025	Fixed	Fixed	No	No	Role Modification
⑬	ADAS [18]	ICLR 2025	Fixed	Fixed	Yes	Yes	Re-Optimization
⑭	AFlow [30]	ICLR 2025	Fixed	Fixed	Yes	Yes	Re-Optimization
⑮	MAV [38]	ICLR-W 2025	Fixed	Fixed	No	No	Yes
⑯	MAS-GPT [19]	ICLR-W 2025	Dynamic	Dynamic	Yes	Yes	Yes
Multi-Agent Systems for Coding Tasks							
⑰	MetaGPT [20]	ICLR 2024	Fixed	Fixed	Yes	No	Coding-Specific
⑱	ChatDev [16]	ACL 2024	Fixed	Fixed	Yes	No	Coding-Specific
⑲	MapCoder [39]	ACL 2024	Fixed	Fixed	Yes	No	Coding-Specific
⑳	EvoMAC [40]	ICLR 2025	Dynamic	Dynamic	Yes	No	Coding-Specific
Multi-Agent Systems for Mathematical Tasks							
㉑	MACM [21]	NeurIPS 2024	Fixed	Fixed	No	No	Math-Specific
Multi-Agent Systems for Scientific Tasks							
㉒	MedAgents [41]	ACL-F 2024	Fixed	Fixed	No	No	Medicine-Specific
Multi-Agent Systems for Data Synthesis							
㉓	MATRIX [25]	ICML 2024	Fixed	Fixed	No	No	Synthesis-Specific

References

[1] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anthropic. Claude 3.5 sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024. Accessed: 2025-01-22.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[4] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[5] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[6] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Method	MATH	GSM-H	AQUA-	AIME-	SciBe	GPQA	GPQA-	MMLU	MedQA	MedMC	Avg-V	Avg-R
Llama-3.3-70B-Instruct												
Single	73.0	55.0	83.9	23.3	40.4	60.7	50.0	85.2	85.6	70.6	63.4	6.4
CoT	74.4	**63.4**	84.2	26.7	38.2	57.6	55.6	84.2	84.4	74.0	64.9	5.1
SC	76.4	54.6	**86.2**	30.0	41.6	62.7	51.5	85.6	86.0	72.6	65.3	4.0
AutoGen	73.0	55.4	83.9	20.0	33.0	59.6	44.4	82.0	82.2	70.2	61.1	8.4
LLM-Debate	78.2	56.0	**86.2**	30.0	**43.6**	**63.2**	53.0	**85.8**	86.4	**75.2**	**66.6**	**2.1**
MAD	76.6	57.8	83.9	**33.3**	42.0	55.6	50.5	83.8	85.2	73.2	64.9	4.6
AgentVerse	79.0	54.0	85.8	23.3	41.0	61.6	53.5	84.8	85.6	73.6	64.9	4.5
DyLAN	77.6	56.2	83.9	**33.3**	41.2	62.5	**56.6**	82.4	**86.6**	73.0	66.0	3.5
MacNet	75.4	59.6	83.5	26.7	37.2	60.9	53.0	82.8	83.8	70.4	63.5	6.7
MAV	**80.0**	38.0	69.3	30.0	36.0	0.5	49.5	0.8	78.2	69.4	47.1	8.3
Qwen-2.5-72B-Instruct												
Single	82.0	68.6	85.8	20.0	42.8	45.1	48.0	83.2	78.8	67.6	62.9	5.5
SC	**85.6**	67.6	**88.6**	20.0	45.0	48.4	50.0	84.0	80.0	69.2	64.7	3.1
AutoGen	81.6	67.8	85.0	13.3	41.6	44.9	45.5	83.8	81.4	68.8	62.3	5.9
LLM-Debate	85.2	66.8	85.4	20.0	44.0	**49.5**	**50.5**	**85.4**	**81.8**	**71.4**	64.9	**2.4**
MAD	83.8	66.4	84.2	20.0	**47.6**	45.8	48.0	81.6	77.6	68.6	62.7	5.7
AgentVerse	82.0	62.6	86.2	13.3	43.6	48.7	**50.5**	83.4	79.0	**71.4**	63.1	4.2
DyLAN	84.4	68.6	88.2	**23.3**	46.4	48.7	47.5	85.0	81.2	70.0	**65.1**	2.7
MacNet	82.0	**69.2**	86.6	10.0	41.6	45.8	44.4	83.8	76.6	65.2	61.1	6.3
MAV	82.4	21.2	53.1	0.0	20.6	0.7	49.5	0.8	74.2	65.8	39.6	8.0
Qwen-2.5-32B-Instruct												
Single	81.8	69.0	87.0	13.3	42.2	45.1	46.5	80.6	72.0	63.6	60.1	4.8
SC	84.8	66.2	**88.6**	16.7	**46.4**	46.4	47.0	81.2	74.8	63.8	61.6	3.1
AutoGen	78.0	**69.4**	85.4	13.3	41.0	45.1	46.0	80.4	72.4	63.0	59.4	5.8
LLM-Debate	84.6	68.6	88.2	**20.0**	43.2	46.4	49.0	**83.0**	75.6	**66.6**	**62.5**	**2.4**
MAD	84.8	68.0	85.8	**20.0**	45.6	45.3	43.4	79.4	71.6	63.6	60.8	4.9
AgentVerse	82.6	64.0	85.4	16.7	41.8	**48.9**	49.5	81.2	74.8	64.8	61.0	4.0
DyLAN	**85.2**	68.2	86.6	13.3	42.2	48.0	**51.0**	81.6	**76.2**	64.6	61.7	2.8
MacNet	81.4	**69.4**	85.8	13.3	39.4	44.9	44.4	80.4	68.8	62.6	59.0	6.3
Qwen-2.5-14B-Instruct												
Single	80.6	**67.0**	83.5	13.3	39.2	42.2	40.9	78.0	70.0	64.2	57.9	4.1
SC	**83.4**	64.4	**85.4**	10.0	42.6	**44.2**	41.4	77.6	69.2	64.4	58.3	3.3
AutoGen	78.6	65.2	84.7	13.3	36.0	42.0	45.5	78.4	68.4	64.0	57.6	4.5
LLM-Debate	81.0	64.0	85.0	**23.3**	42.2	42.0	**47.0**	**80.2**	71.0	65.6	**60.1**	2.6
MAD	79.0	61.8	82.7	6.7	**44.4**	37.5	41.4	75.2	65.0	59.0	55.3	6.2
AgentVerse	61.2	38.4	62.6	16.7	36.4	42.9	41.9	47.0	61.2	46.8	45.5	6.4
DyLAN	83.0	64.4	85.0	13.3	41.8	43.3	46.5	78.8	**72.0**	**66.2**	59.4	**2.3**
MacNet	78.2	66.2	83.1	13.3	38.4	41.5	41.4	75.4	68.0	59.4	56.5	5.4
Qwen-2.5-7B-Instruct												
Single	77.0	60.6	78.3	**16.7**	32.8	**37.0**	33.3	**74.2**	63.0	56.0	52.9	3.9
SC	**80.6**	57.2	**85.0**	**16.7**	**35.8**	35.9	36.9	72.4	63.2	57.4	54.1	3.2
AutoGen	74.2	57.4	81.9	6.7	31.6	32.8	38.9	72.4	63.0	54.4	51.3	5.7
LLM-Debate	80.0	**61.4**	84.2	13.3	34.6	34.6	38.9	**74.2**	**65.0**	57.8	**54.4**	**2.6**
MAD	71.4	59.6	79.1	**16.7**	33.2	27.9	28.3	55.0	36.8	32.6	44.1	6.6
AgentVerse	76.2	59.4	82.3	13.3	34.4	35.7	**40.9**	72.2	59.2	58.2	53.2	4.4
DyLAN	80.0	60.2	80.3	10.0	33.6	**37.0**	40.4	71.8	60.8	**59.4**	53.4	3.8
MacNet	76.6	60.0	79.9	**16.7**	30.0	36.4	38.9	72.6	62.4	57.2	53.1	4.3

Table 2: Main results. Avg-V denotes averaged performance value across benchmarks (higher is better) while Avg-R denotes averaged rank across benchmarks (lower is better).

[7] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[8] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[9] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

[10] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68, 2024.

[11] Yotam Wolf, Noam Wies, Oshri Avnery, Yoav Levine, and Amnon Shashua. Fundamental limitations of alignment in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53079–53112, 2024.

[12] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[13] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, 2023.

[14] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.

[15] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2023.

[16] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, 2024.

[17] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[18] Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

[19] Rui Ye, Shuo Tang, Rui Ge, Yaxin Du, Zhenfei Yin, Jing Shao, and Siheng Chen. MAS-GPT: Training LLMs to build LLM-based multi-agent systems. In *Workshop on Reasoning and Planning for Large Language Models*, 2025.

[20] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2024.

[21] Bin Lei, Yi Zhang, Shan Zuo, Ali Payani, and Caiwen Ding. Macm: Utilizing a multi-agent system for condition mining in solving complex mathematical problems. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[22] Shima Imani, Liang Du, and Harsh Shrivastava. Mathprompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, 2023.

[23] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

[24] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025.

[25] Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. Self-alignment of large language models via monopolylogue-based social scene simulation. In *International Conference on Machine Learning*, pages 39416–39447. PMLR, 2024.

[26] Shuo Tang, Xianghe Pang, Zexi Liu, Bohan Tang, Rui Ye, Tian Jin, Xiaowen Dong, Yanfeng Wang, and Siheng Chen. Synthesizing post-training data for llms through multi-agent simulation. *arXiv preprint arXiv:2410.14251*, 2024.

[27] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2024.

[28] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[29] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

[31] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic llm-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*, 2024.

[32] Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. Gptswarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*, 2024.

[33] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[34] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[35] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2024.

[36] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

[37] Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Scaling large language model-based multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations*, 2025.

[38] Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with goal verifiers. In *Workshop on Reasoning and Planning for Large Language Models*, 2025.

[39] Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. Mapcoder: Multi-agent code generation for competitive problem solving. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4912–4944, 2024.

[40] Yue Hu, Yuzhu Cai, Yaxin Du, Xinyu Zhu, Xiangrui Liu, Zijie Yu, Yuchen Hou, Shuo Tang, and Siheng Chen. Self-evolving multi-agent networks for software development. In *The Thirteenth International Conference on Learning Representations*, 2025.

[41] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 599–621, 2024.