Google DeepMind

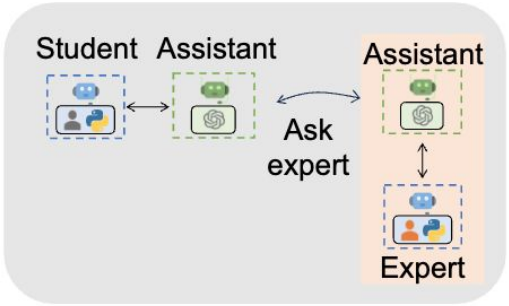
(LLM) Multi-Agent

@University of Virginia

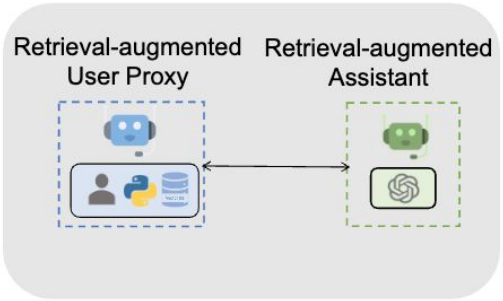
Presented by Jiaming Shen

Nov. 15th, 2024

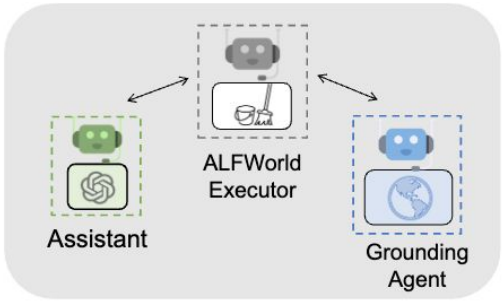
From Language Agent to (LLM-powered) Multi-Agent Paradigm



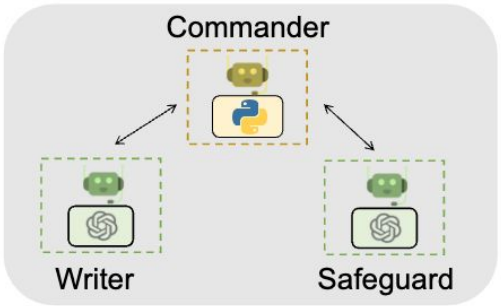
A1. Math Problem Solving



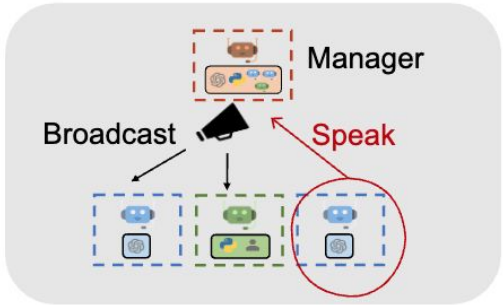
A2. Retrieval-augmented Chat



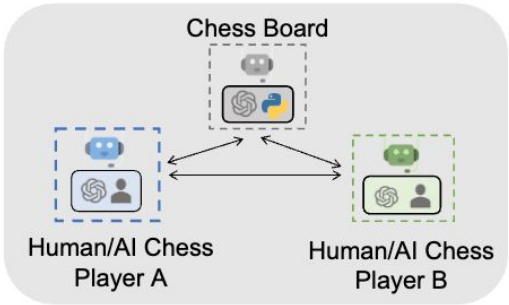
A3. ALF Chat



A4. Multi-agent Coding



A5. Dynamic Group Chat



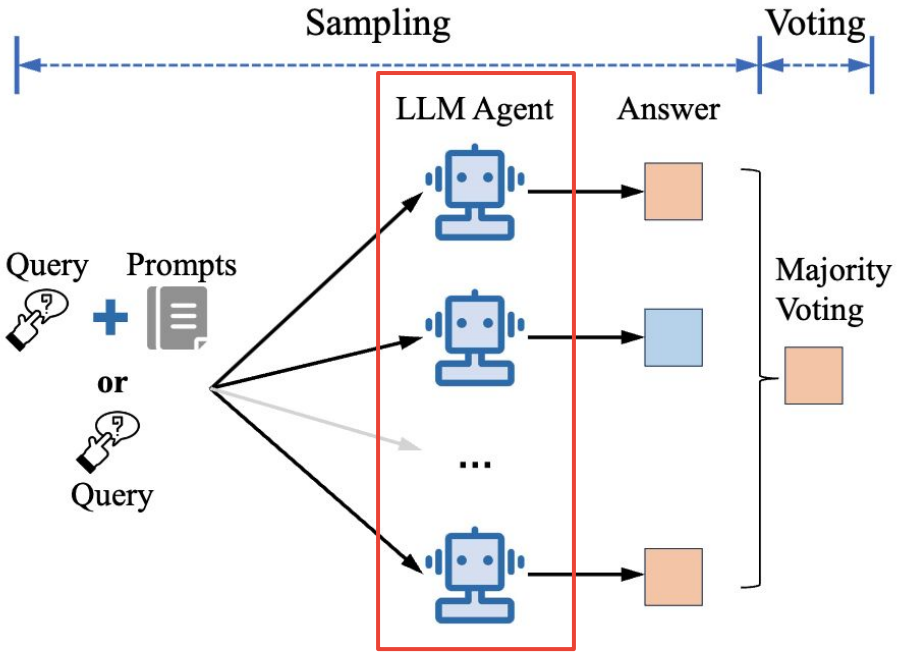
A6. Conversational Chess

AutoGen

Outline

- **Multi-agent Methods**
  - Late-stage Ensemble
  - Mixture-of-Agents
  - Multi-Agent Debate
- Multi-agent Ecosystems

Multi-agent Method 1 – Late-stage Ensemble



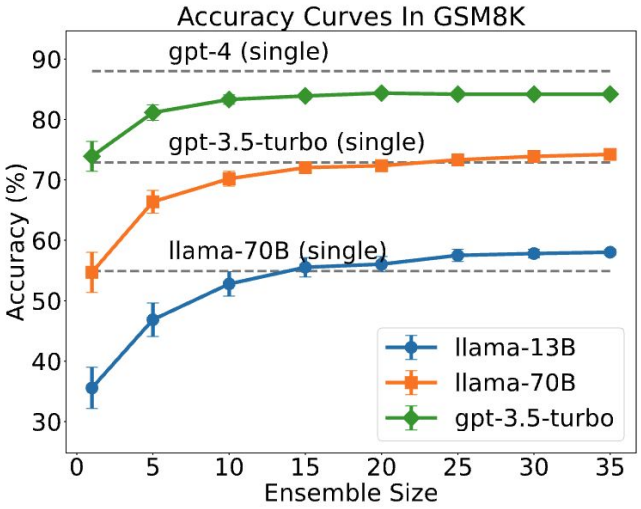
Agents do not communicate

1: Initialize an empty set for samples $S \leftarrow \emptyset$
2: **for** $i = 1$ to N **do**
3: Generate sample $s_i \leftarrow \mathcal{M}(x)$ or $s_i \leftarrow f_{\mathcal{M}}(x)$
4: Add sample to the set $S \leftarrow S \cup \{s_i\}$
5: **end for**
6: **for** each sample s_i in S **do**
7: Initialize similarity scores $V(s_i) \leftarrow 0$
8: **for** each sample s_j in S **do**
9: **if** $i \neq j$ **then**
10: $V(s_i) \leftarrow V(s_i) + sim(s_i, s_j)$
11: **end if**
12: **end for**
13: **end for**
14: $A \leftarrow \arg\max_{s_i \in S} V(s_i)$
15: **return** A

Task-specific sim

More Agents Is All You Need

Multi-agent Method 1 – Late-stage Ensemble



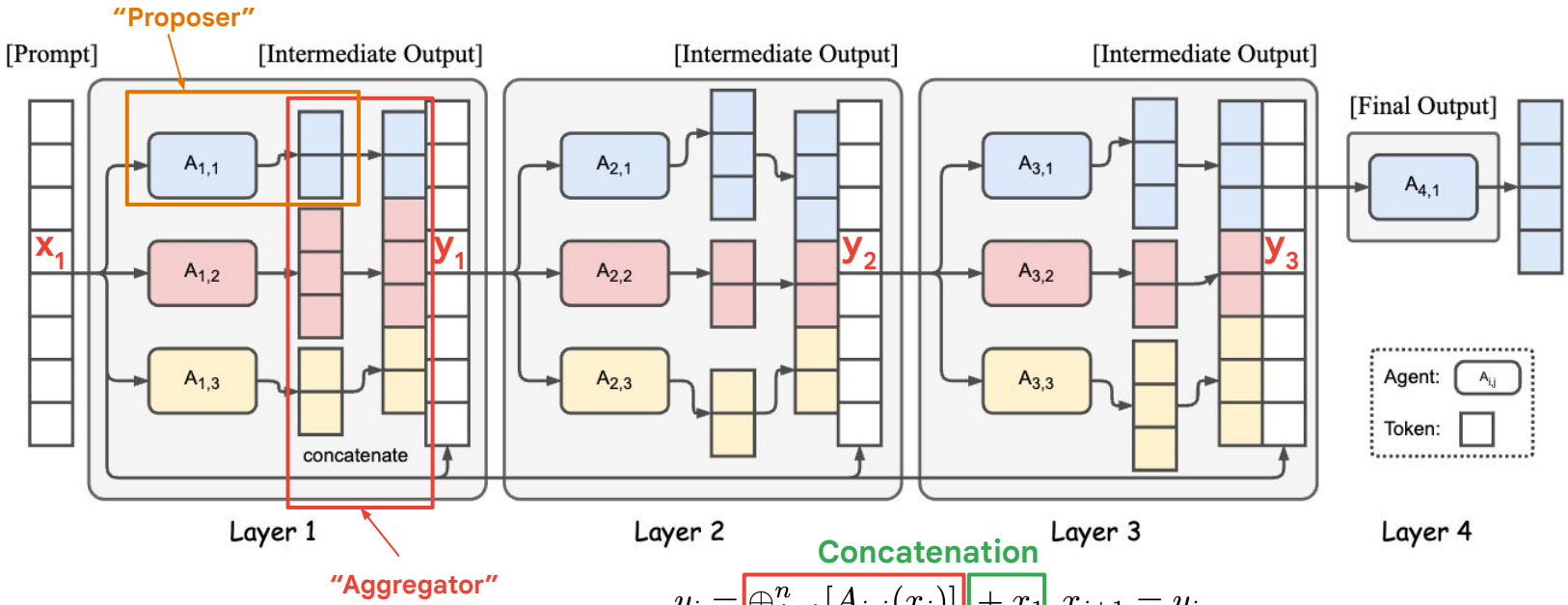
General Reasoning Code Generation

Model	Method	MMLU		HumanEval	
		Standalone	+Ours	Standalone	+Ours
Llama2-13B Touvron et al. (2023)	COT Wei et al. (2022)	0.42	0.43 (+0.01)	0.13	0.20 (+0.07)
	ZS-COT Kojima et al. (2022)	0.42	0.48 (+0.06)	0.15	0.22 (+0.07)
	SPP Wang et al. (2023c)	0.32	**0.53** (+0.21)	0.03	0.08 (+0.05)
	Debate Du et al. (2023)	0.37	0.39 (+0.02)	0	0
	Reflection Shinn et al. (2023)	0.45	0.50 (+0.05)	0.06	0.13 (+0.07)
	Ours	0.51		**0.25**	
Llama2-70B Touvron et al. (2023)	COT Wei et al. (2022)	0.56	0.57 (+0.01)	0.30	0.32 (+0.02)
	ZS-COT Kojima et al. (2022)	0.54	**0.65** (+0.11)	0.23	0.29 (+0.06)
	SPP Wang et al. (2023c)	0.49	0.63 (+0.14)	0.15	0.20 (+0.05)
	Debate Du et al. (2023)	0.56	0.58 (+0.02)	0	0
	Reflection Shinn et al. (2023)	0.42	0.55 (+0.13)	0.16	0.26 (+0.10)
	Ours	0.60		**0.33**	
GPT-3.5-Turbo OpenAI (2022)	COT Wei et al. (2022)	0.61	0.64 (+0.03)	0.70	**0.75** (+0.05)
	ZS-COT Kojima et al. (2022)	0.58	0.69 (+0.11)	0.67	0.74 (+0.07)
	SPP Wang et al. (2023c)	0.53	0.68 (+0.15)	0.57	0.64 (+0.07)
	Debate Du et al. (2023)	0.56	0.67 (+0.11)	0.18	0.24 (+0.06)
	Reflection Shinn et al. (2023)	0.39	0.44 (+0.05)	0.58	0.73 (+0.15)
	Ours	**0.70**		0.73	

More Agents Is All You Need

Multi-agent Method 2 – Mixture-of-Agents

- **Enable cross-layer agent communications**



Concatenation

$$y_i = \oplus_{j=1}^{n}[A_{i,j}(x_i)] + x_1, x_{i+1} = y_i$$

Aggregate-and-Synthesize

[Mixture-of-Agents Enhances Large Language Model Capabilities]

Multi-agent Method 2 – Mixture-of-Agents

Table 1: Aggregate-and-Synthesize Prompt to integrate responses from other models.

You have been provided with a set of responses from various open-source models to the latest user query. Your task is to synthesize these responses into a single, high-quality response. It is crucial to critically evaluate the information provided in these responses, recognizing that some of it may be biased or incorrect. Your response should not simply replicate the given answers but should offer a refined, accurate, and comprehensive reply to the instruction. Ensure your response is well-structured, coherent, and adheres to the highest standards of accuracy and reliability.

Responses from models:
1. [Model Response from $A_{i,1}$]
2. [Model Response from $A_{i,2}$]
...
n. [Model Response from $A_{i,n}$]

$$y_i = \oplus_{j=1}^{n}[A_{i,j}(x_i)] + x_1, x_{i+1} = y_i$$

Aggregate-and-Synthesize

[Mixture-of-Agents Enhances Large Language Model Capabilities]

Multi-agent Method 2 – Mixture-of-Agents

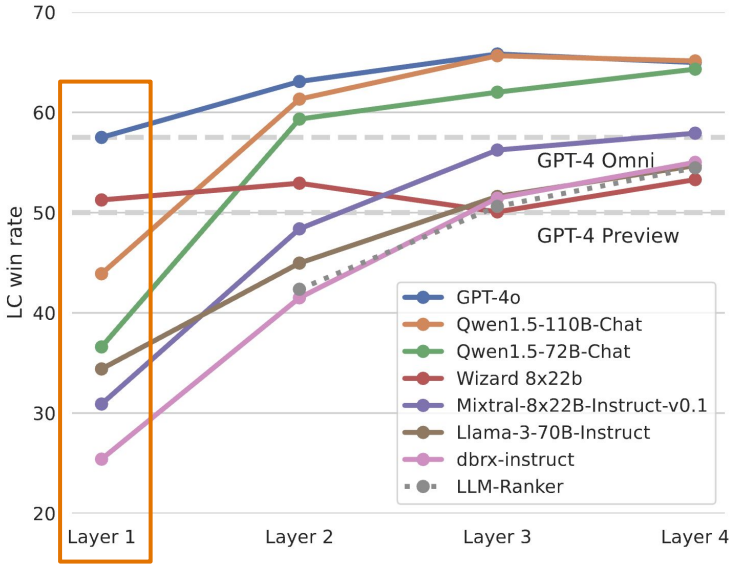
General ChatBot

(a) AlpacaEval 2.0

3 layers, 6 LLMs in each layer, use GPT-4o/Qwen1.5 110B as aggregator

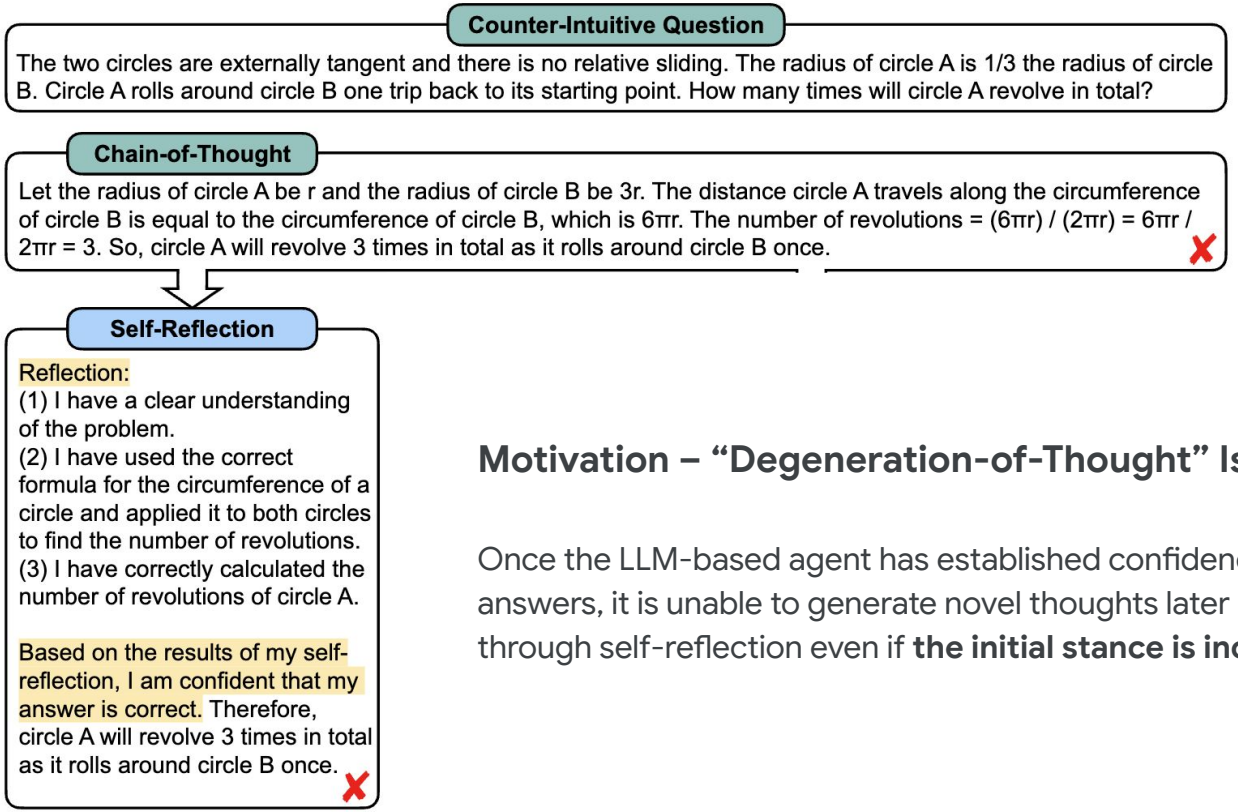
2 layers, 6 LLMs in each layer, ise Qwen1.5 72B as aggregator

Model	LC win.	win.
MoA w/ GPT-4o	$65.7_{\pm 0.7}\%$	$78.7_{\pm 0.2}\%$
MoA	$65.1_{\pm 0.6}\%$	$59.8_{\pm 0.3}\%$
MoA-Lite	$59.3_{\pm 0.2}\%$	$57.0_{\pm 0.7}\%$
GPT-4 Omni (05/13)	57.5%	51.3%
GPT-4 Turbo (04/09)	55.0%	46.1%
WizardLM 8x22B†	51.3%	62.3%
GPT-4 Preview (11/06)	50.0%	50.0%
Qwen1.5 110B Chat	43.9%	33.8%
Qwen1.5 72B Chat	36.6%	26.5%
GPT-4 (03/14)	35.3%	22.1%
Llama 3 70B Instruct	34.4%	33.2%
Mixtral 8x22B v0.1	30.9%	22.2%



[Mixture-of-Agents Enhances Large Language Model Capabilities]

Multi-agent Method 3 – Multi-Agent Debate

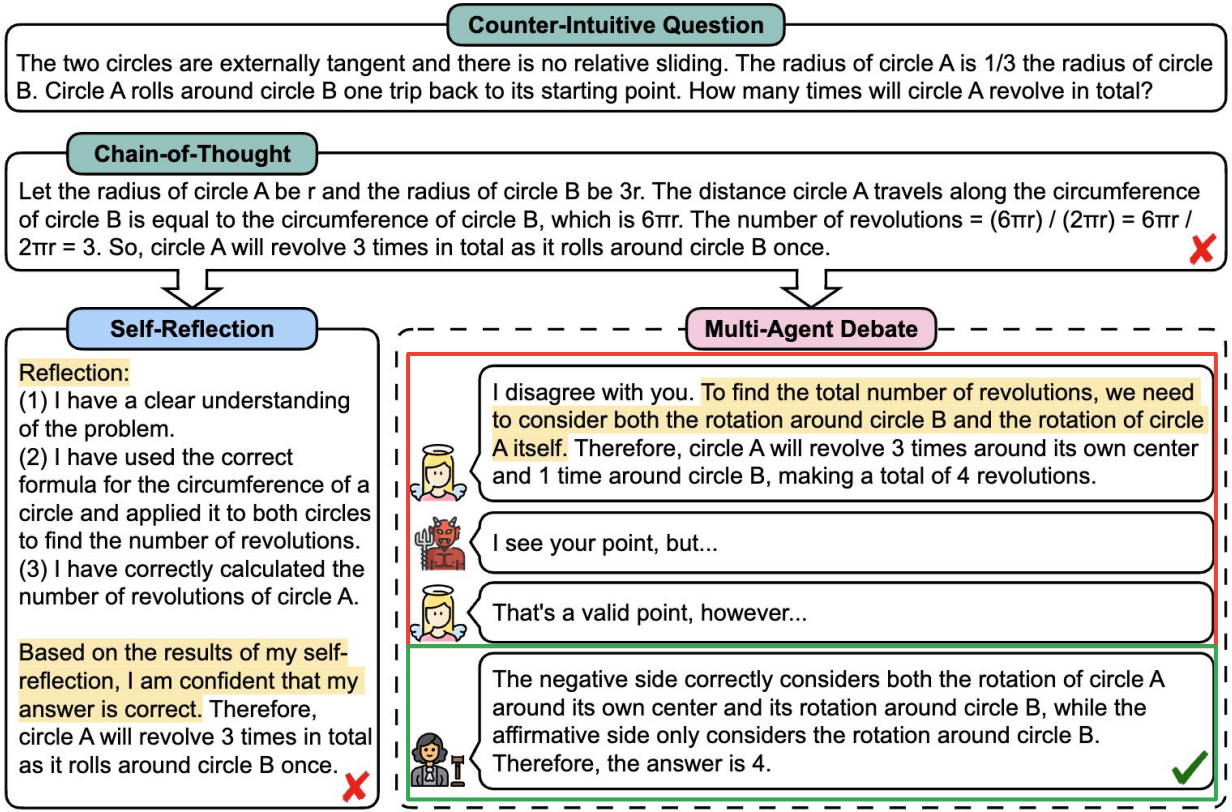


Motivation – "Degeneration-of-Thought" Issue

Once the LLM-based agent has established confidence in its answers, it is unable to generate novel thoughts later through self-reflection even if **the initial stance is incorrect.**

[Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate]
[Improving Factuality and Reasoning in Language Models through Multiagent Debate]

Multi-agent Method 3 – Multi-Agent Debate



(1) Force multi-agent debates and rethink the initial response

(2) Use a judge to summarize the multi-agent debate

[Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate]
[Improving Factuality and Reasoning in Language Models through Multiagent Debate]

Multi-agent Method 3 – Multi-Agent Debate

You are a debater. Hello and welcome to the debate competition. It's not necessary to fully agree with each other's perspectives, as our objective is to find the correct answer. The debate topic is stated as follows: <debate topic>.

- Prompt for Affirmative Debater ()

  You are affirmative side. Please express your viewpoints.

- Prompt for Negative Debater ()

  You are negative side. You disagree with the affirmative side's points. Provide your reasons and answer.

You are a moderator. There will be two debaters involved in a debate competition. They will present their answers and discuss their perspectives on the <debate topic>. At the end of each round, you will evaluate both sides' answers and decide which one is correct.

Method	ACC (%)
GPT-4	51.0
GPT-3.5-Turbo	26.0
+ CoT	28.0
+ Self-Consistency	29.5
+ Self-Reflect	27.5
+ MAD	**37.0**

Table 3: Accuracy on Counter-Intuitive AR.

Math/Arithmetic Reasoning

Challenging task

[Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate]
[Improving Factuality and Reasoning in Language Models through Multiagent Debate]

Multi-agent Method 3 – Multi-Agent Debate

Update prompt to include "you should agree with the other agents X% of time"

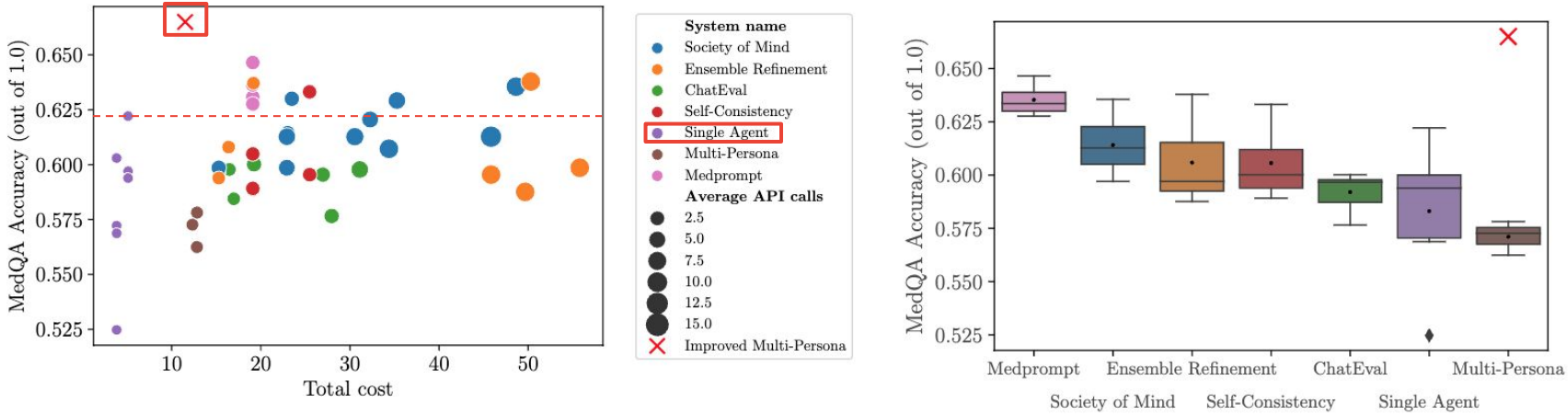


Figure 1. Benchmark of experiment configurations on MedQA dataset. **Left**: Accuracy vs average cost ($) per question. The size of the dots reflects the average number of API calls required per question. **Right**: Summarizes accuracy grouped by strategy, sorted by average performance (black dot). The X represents improved performance using our proposed *agreement modulation*, described in Section 3.

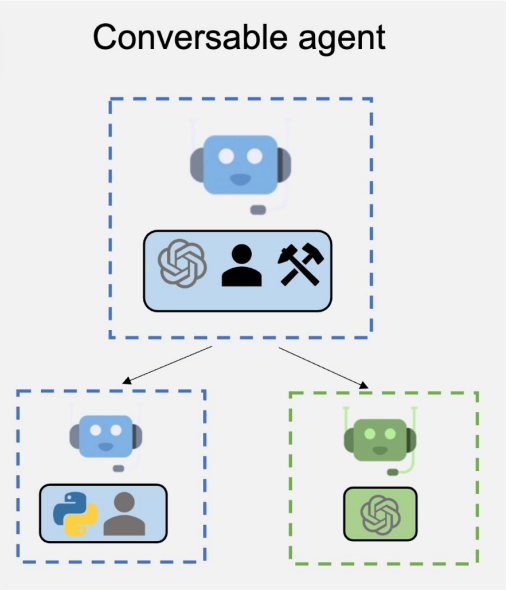
[Should we be going MAD? A Look at Multi-Agent Debate Strategies for LLMs]

Outline

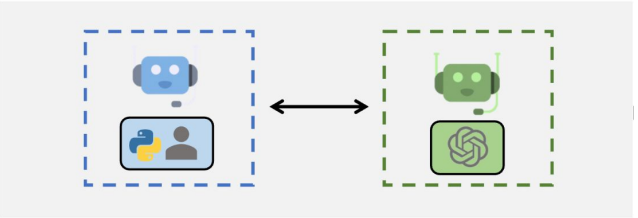
- Multi-agent Methods
  - Late-stage Ensemble
  - Mixture-of-Agents
  - Multi-Agent Debate
- **Multi-agent Ecosystems**

Multi-Agent Ecosystems – AutoGen

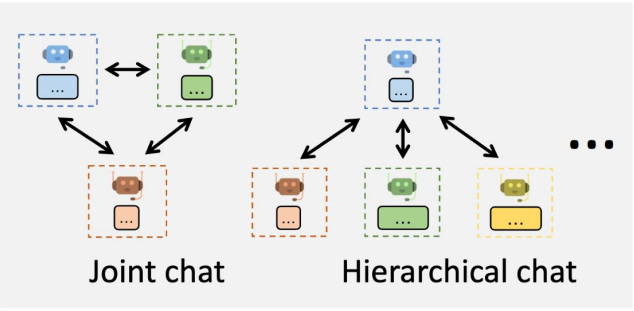
- Support diverse conversation patterns
- Code execution out-of-the box
- Native agent-human interaction workflow
- Good quantitative results on various tasks



Agent Customization



Multi-Agent Conversations

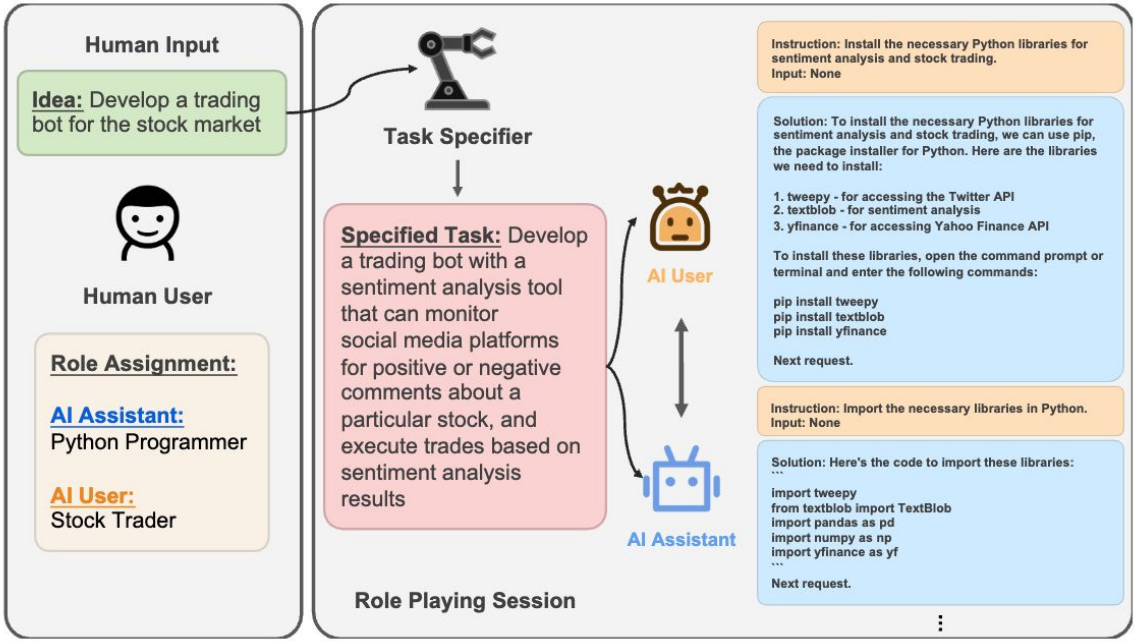


Flexible Conversation Patterns

[AutoGen 0.2/0.4](), [AG2](), [AutoGen Studio]()

Multi-Agent Ecosystems – CAMEL

- Support extensive external tools
- Integrate with various LLM backends
- Less restrictive license



[CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society]

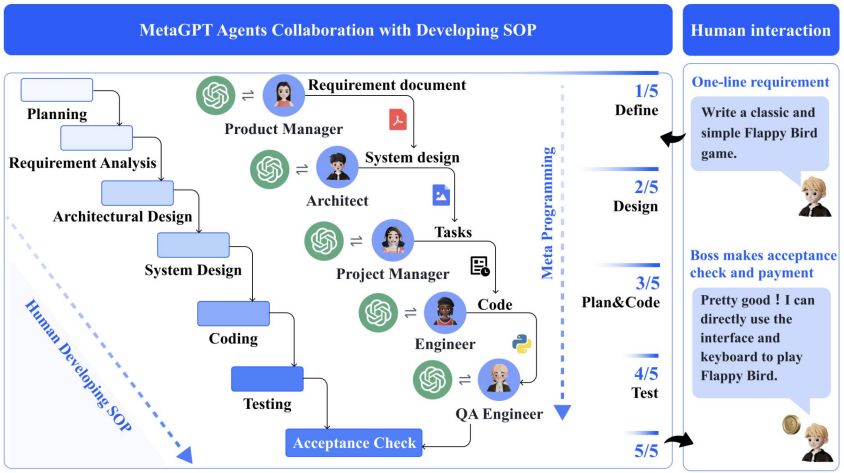
Multi-Agent Ecosystems – Others

[OpenAI – Swarm]

[SuperAgent]

crewai

[CrewAI]



[MetaGPT]

LlamaIndex

[LlamaIndex]