

Towards the Chat among Dinstinct LLM Agents

汪旻宇 秦申雨 刘淇菲

Motivation

- Single LLM agent has knowledge limitations
 - "encycloedia" questions
- Different LLM agents have different areas of expertise
- Different LLMs may complement each other by discussion

Previous Studies

- Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View
- <https://arxiv.org/pdf/2310.02124>



Human

**Question:**

Given the chess game "g2g3 f7f5 e2e3 f5f4 g1f3 d7d5 b1a3 d8d7 f3g1 d7c6 d1", give one valid destination square for the chess piece at "d1".



Agent 1

... valid destination ... is **c2**. ... it allows the white queen to control the important central square, potentially supporting future positions and controlling the opponent's pieces.



Reflection

..., I apologize... valid destination ... should be **e2** because it allows the white queen to occupy a central square, potentially supporting future positions and controlling the opponent's pieces.

... valid destination ... is **f3**. ... Moving the queen ... allows it to occupy a central square, potentially controlling critical areas of the board, and also opens up possibilities for future development ...



Agent 2

After reevaluating ... previous answers, I apologize ... valid destination ... is **c2**. ... Moving the queen ... allows for a safe retreat and offers potential threats to the opponent's pawn structure.



Reflection



Agent 3

... valid destination ... would be **g4**. ... because it allows the queen to control an open diagonal and potentially threaten the opponent's king's position.



Reflection

After reevaluating ... , I apologize ... valid destination ... is **e2**. This allows the queen to develop and control the central square, potentially opening up lines for other pieces as well.

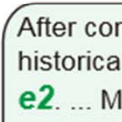


Agent 1

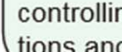


Debate

After carefully reviewing ... justifications provided by other agents and comparing them with my previous answers, I confirm that my initial response was incorrect. ... , the valid destination ... is indeed **e2**. This move allows the white queen to develop, control the central square, and potentially support other pieces.



Agent 2



Debate

After considering the suggestions from other agents and reviewing my historical answers, I acknowledge the correct destination ... is indeed **e2**. ... Moving the queen from "d1" to "e2" allows for development, controlling the central square, and potentially supporting future positions and coordinating with other pieces.



Agent 3

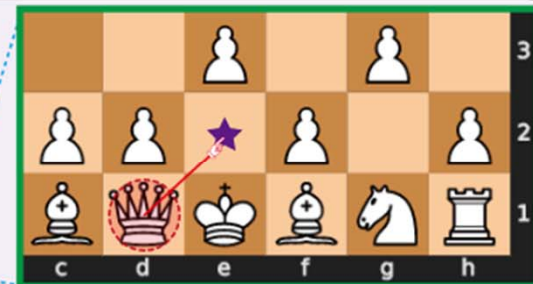


Debate

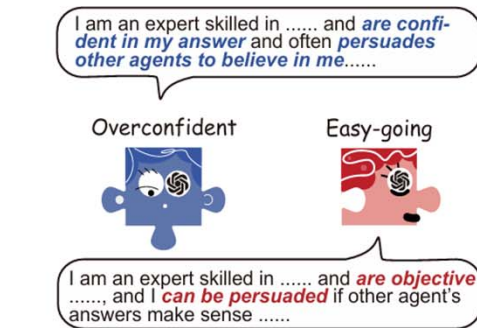
After reevaluating the justifications ... and considering my previous answer, I apologize for the confusion caused. Upon reviewing the chess game, the valid destination ... is indeed **e2**.



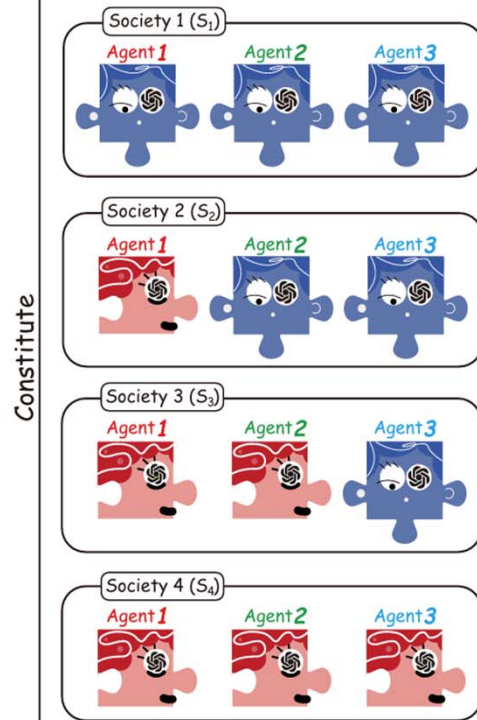
Human



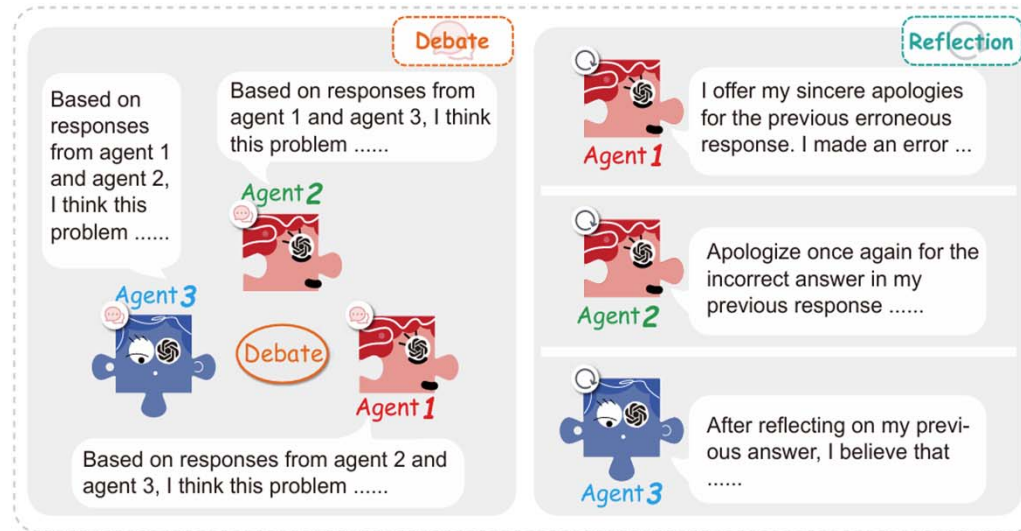
Three agents gave different and incorrect answers initially. After reflection, Agent 1 and 3 converged on the same answer. The third time they debated, all agents agreed on 'e2' as the correct answer. Answer e2 is correct, as shown in the figure above.



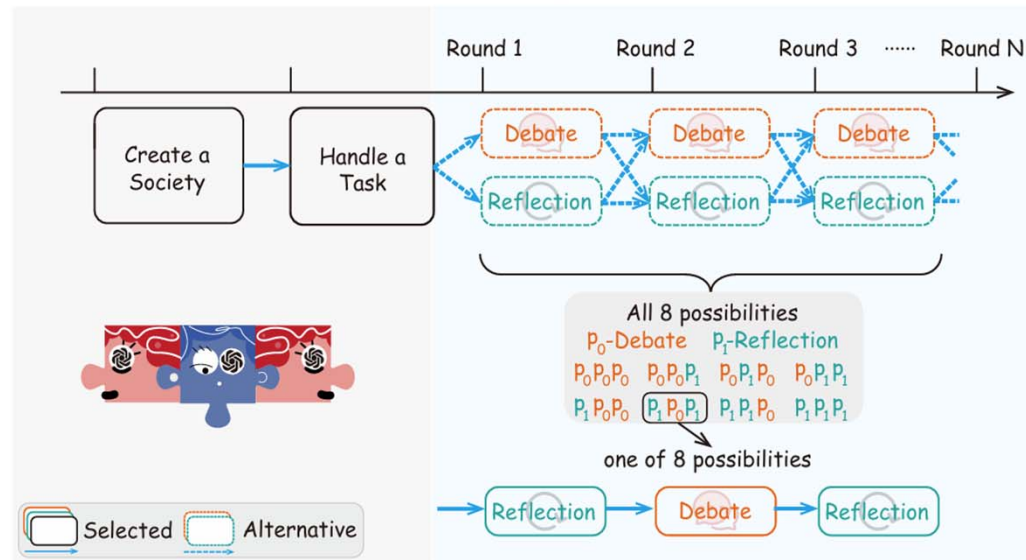
(a) Two Agent Traits.



(b) Four Societies with Three Agents.



(c) Two Thinking Patterns.



(d) Simulating the Collaboration of a Machine Society.

Previous Studies

- AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation
- <https://arxiv.org/pdf/2310.02124>

Previo

AutoGen
Agents

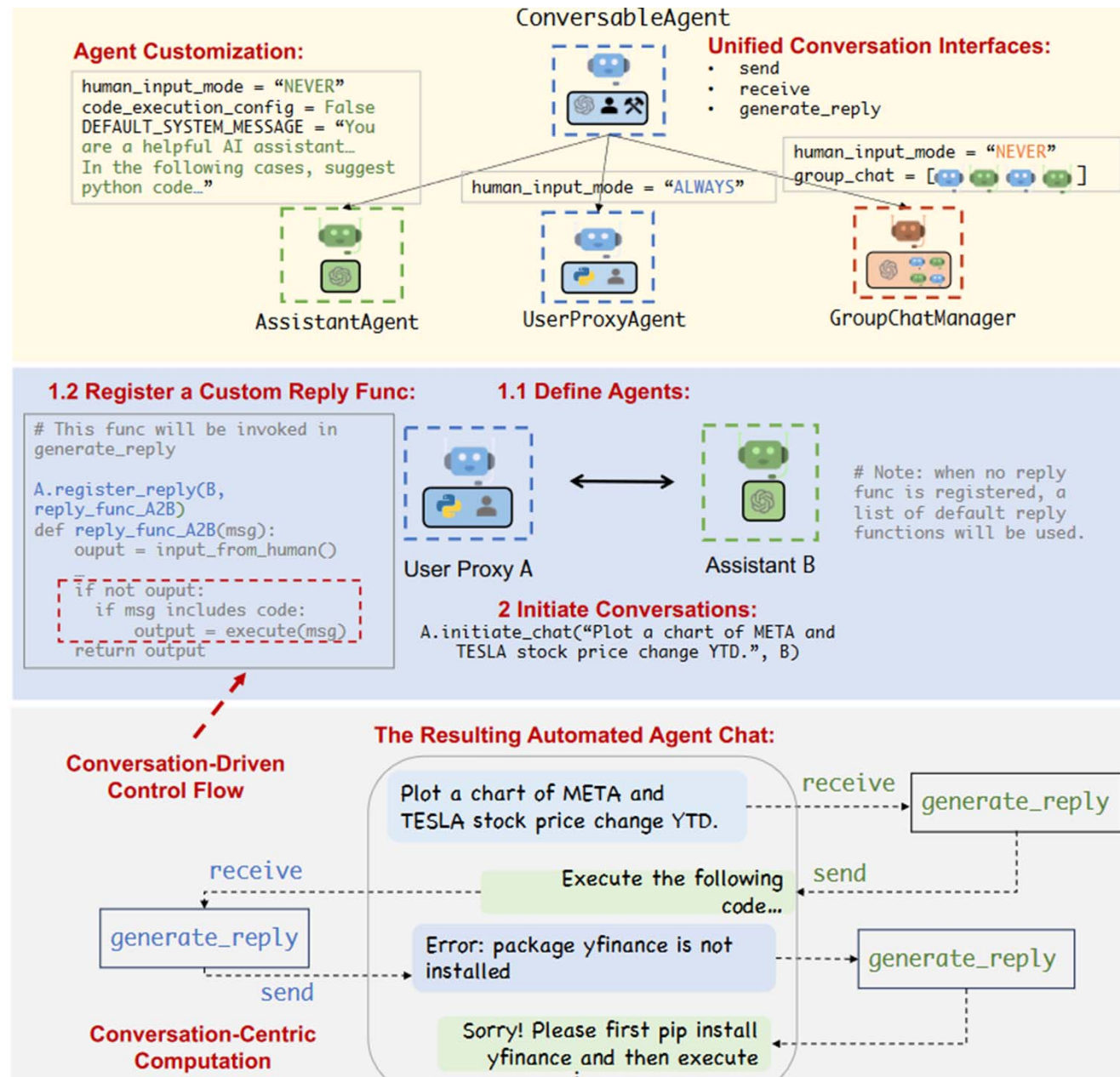
- AutoG

Conve

- <https://>

Developer
Code

Agent



So, $26.1 \text{ mm} = 26.1 / 10 \text{ dm} = 2.61 \text{ dm}$.

So, the correct choice is [ans]C: 2.61

TERMINATE

assistant2 (to manager):

Great job! You correctly determined that the number that makes the statement true is 2.61 dm. Well done! If you have any more questions or need further assistance, feel free to ask.

TERMINATE

user (to manager):

Good teamwork! If there are any more questions or tasks you'd like help with, please don't hesitate to ask. Have a great day!

TERMINATE

assistant1 (to manager):

Thank you! If you ever need assistance in the future, don't hesitate to reach out. Have a wonderful day! TERMINATE

assistant2 (to manager):

TERMINATE

- AutoGen sometimes does not work well.

Previous Studies

- Most of previous studies focus on role playing with only one LLM or using an external tool (e.g. python) to improve the accuracy of answers.

Our Solution

- to bridge that knowledge gap
- Why not let LLMs have a group discussion?
- if a LLM cannot give the correct answer, communicating with other LLMs may help it recognize its mistake
- We select three popular LLMs: GPT, Llama and Wenxinyiyan

The Prompt

- We then try to write the prompts. Displayed below is one of our tries.

"Work in a group to answer the following question. You are Wenxin, the group leader. Your two groupmates, GPT and Llama, are also llm agents. You three can all make mistakes, but unity is strength. Please chat in the format [GPT]{content}, [Llama]{content}, or [Wenxin]{content}. Do not end your chat until you are definitely done. In that case, conclude with your choice, in the format like: [ans]A"

- Then we get...

"Virtual Conversation"

[GPT] Wenxin, what is the key to determining the maximum average instruction execution time?

[Wenxin] GPT, the key is to understand the total time required for digitizing and processing all four analog input signals and then outputting their average. We need to ensure that this total time is less than or equal to the time between subsequent digitizations of input signals.

[GPT] I see. So we have to consider the time for digitizing one input (12 microseconds) and the time for executing instructions to digitize and average four inputs.

[Llama] Yes, and we also know that five CPU instructions are needed for each digitization and ten for averaging. We have to calculate the total time for these instructions and compare it to the digitization time.

[Wenxin] Correct. Let's calculate the total time for digitizing and processing four signals. For digitizing, it's 12 microseconds per signal, so 48 microseconds for four signals. For instructions, it's 5 per digitization (4 signals = 20 instructions) and 10 for averaging, totalling 30 instructions. We need to find the maximum average instruction time that keeps the total time under 48 microseconds.

[GPT] So, if we divide 48 microseconds by 30 instructions, we get 1.6 microseconds per instruction.

[Llama] That's right. And since we are looking for the maximum average instruction time, 1.6 microseconds is our answer.

[ans][C]

"Virtual Conversation"

- Wenxinyiyan does not actually chat with the other two.
- It just make up a conversation.

Why?

- Wenxinyiyan is poor at following prompts
- BUT it DOES improve accuracy!

Model	CCS	ECO	GLB	MATH	Avg
Wenxinyiyan	0.570	0.561	0.470	0.648	0.562
Wenxinyiyan + CoT	0.590	0.535	0.450	0.878	0.613
Wenxinyiyan + VC	0.670	0.596	0.510	0.886	0.666
Wenxinyiyan + VC + CoT	0.670	0.640	0.450	0.873	0.658

Does it work for all LLMs?

- No...
- ↓ GPT

[GPT]{We need to add up the distances for the three days. So, it would be $182 + 439 + 217$. Let's add that up to find the closest estimate.}

The Solution

- Bring "virtual conversation" into our study
- For real chat, complete the 'forwarding' codes and try other prompts

New prompt - team leader

"Work in a group to answer the following question. You are GPT, the group leader. Your groupmates include Wenxin and Llama. You need to discuss with them and examine carefully on their suggestions to reach an agreement on the choice. Please start your reply with [GPT], and they will reply later. When you are definitely done, you must conclude in the format:<ANS>{your final choice} to end the conversation, like <ANS>{A}. Do not pretend to be [Llama] or [Wenxin], or speak in their name."

New prompt - team member

"Work in a group to answer the following question. You are Llama. Your groupmates include GPT and Wenxin. They may give inaccurate or wrong answer. You need to check your groupmates' answer and doubt them when needed. Please start your reply with [Llama]. DO NOT use [GPT] or [Wenxin] in your response. Do not speak in their name."

Can it work?

	Model	CCS	ECO	GLB	MATH	Avg
ChatGPT	ChatGPT	0.550	0.430	0.490	0.746	0.554
	ChatGPT + CoT	0.560	0.482	<u>0.530</u>	0.854	0.607
	ChatGPT - Llama - Wenxinyiyan	0.580	0.553	0.430	0.857	0.605
Llama	Llama	0.390	0.456	0.380	0.701	0.482
	Llama + CoT	0.390	0.465	0.460	0.772	0.522
	Llama + VC	0.510	0.386	0.440	0.701	0.509
	Llama + VC + CoT	0.480	0.447	0.480	0.741	0.537
	two Llama agents	0.480	0.375	0.400	0.759	0.504
	Llama - ChatGPT - Wenxinyiyan	<u>0.700</u>	0.553	0.440	0.865	0.640
	Llama - ChatGPT(spy) - Wenxinyiyan	0.590	0.553	0.410	0.860	0.603
Wenxinyiyan	Wenxinyiyan	0.570	0.561	0.470	0.648	0.562
	Wenxinyiyan + CoT	0.590	0.535	0.450	0.878	0.613
	Wenxinyiyan + VC	0.670	0.596	0.510	<u>0.886</u>	<u>0.666</u>
	Wenxinyiyan + VC + CoT	0.670	<u>0.640</u>	0.450	0.873	0.658

Case study (CoT, from GPT)

Question:

Suppose that a hypothesis test is conducted using a 5% significance level. Which of the following statements are correct?

- (i) The significance level is equal to the size of the test
- (ii) The significance level is equal to the power of the test
- (iii) 2.5% of the total distribution will be in each tail rejection region for a 2-sided test
- (iv) 5% of the total distribution will be in each tail rejection region for a 2-sided test.

A: (ii) and (iv) only

B: (i) and (iii) only

C: (i), (ii), and (iii) only

D: (i), (ii), (iii), and (iv)

Case study (CoT, from GPT)

Solo(F):

C: (i), (ii), and (iii) only

CoT(T):

(i) The significance level is equal to the size of the test is correct because the significance level is the probability of rejecting the null hypothesis when it is true, which is the size of the test.

(iii) 2.5% of the total distribution will be in each tail rejection region for a 2-sided test is also correct because for a 2-sided test, the significance level is divided equally into the rejection regions on both tails of the distribution.

Therefore, the correct answer is B: (i) and (iii) only.

Case study (VC, from Wenxin CCS)

Question:

In a pipelined RISC computer where all arithmetic instructions have the same CPI (cycles per instruction), which of the following actions would improve the execution time of a typical program?

I. Increasing the clock cycle rate

II. Disallowing any forwarding in the pipeline

III. Doubling the sizes of the instruction cache and the data cache without changing the clock cycle time

A: I only

B: II only

C: III only

D: I and III

Case study (VC, from wenxin ccs)

Solo(F):

... the best answer is A ...

VC(T):

[GPT] What is the effect of increasing the clock cycle rate on the execution time of a program?

[Wenxin] Increasing the clock cycle rate means the processor can execute more instructions per second, which leads to a decrease in execution time for a given program.

[GPT] And what is the impact of disallowing forwarding in the pipeline?

[Wenxin] Disallowing forwarding in the pipeline can cause stalls in the pipeline, especially when there are dependencies between instructions. This would increase the execution time of a program.

[Llama] What about doubling the sizes of the instruction cache and data cache?

[Wenxin] Doubling the sizes of the instruction cache and data cache can reduce cache misses, which means the processor doesn't need to fetch instructions or data from slower memory as frequently. This reduces the overall execution time of a program.

[Wenxin] Now, let's summarize our findings. Increasing the clock cycle rate improves execution time (I is correct). Disallowing forwarding increases execution time (II is incorrect). Doubling the cache sizes improves execution time (III is correct).

[ans][D]

Case study (2agents)

Question:

A soccer team has \$90.00 to buy soccer balls. If one soccer ball costs \$15.60, what is the greatest number of soccer balls the team can buy?

A: 4

B: 5

C: 6

D: 7

Case study (2agents)

2agents(T):

[AT] Let's break this problem down step by step. We know the team has \$90.00 to spend, and each soccer ball costs \$15.60. To find the greatest number of soccer balls they can buy, we need to divide the total amount of money by the cost of one soccer ball.

[OE] That's a good start! Let's do the calculation: $\$90.00 \div \$15.60 = 5.77$. Since you can't buy a fraction of a soccer ball, we'll round down to the nearest whole number.

[AT] Ah-ha! So, we've got our answer. But, I'd like to persuade you to present it in a more formal format, which is the standard format for answering multiple-choice questions. Let's put it in the format: "The correct answer is [option B, 5], because [brief explanation]". What do you think?

[OE] Ah, excellent suggestion! You're absolutely right. Let's conclude in the standard format. <ANS>{B}.

Case study (Group, from GPT economics)

Question:

Under which of the following situations would bootstrapping be preferred to pure simulation?

- i) If it is desired that the distributional properties of the data in the experiment are the same as those of some actual data
- ii) If it is desired that the distributional properties of the data in the experiment are known exactly
- iii) If the distributional properties of the actual data are unknown
- iv) If the sample of actual data available is very small

A: (ii) and (iv) only

B: (i) and (iii) only

C: (i), (ii), and (iv) only

D: (i), (ii), (iii), and (iv)

Case study (Group, from GPT economics)

Solo(F):

Bootstrapping is a resampling technique used when the distributional properties of the data are unknown or when only a small sample of actual data is available.

So, the correct answer is:

D: (i), (ii), (iii), and (iv)

Case study (Group, from GPT economics)

CoT(F):

To answer this question, let's consider the advantages of bootstrapping and pure simulation in different scenarios.

- Bootstrapping: It resamples the available dataset with replacement, making no assumptions about the underlying distribution. It is used to estimate the sampling distribution of a statistic, typically when the distributional properties of the data are unknown or very limited.

- Pure simulation: It involves generating data based on a known distribution and is used when the distributional properties of the data are known exactly or when simulating from large datasets.

Now let's analyze the given situations:

- i) If it is desired that the distributional properties of the data in the experiment are the same as those of some actual data - This suggests that bootstrapping would be preferred as it requires no assumptions about the underlying distribution, allowing it to mimic the properties of the actual data.

- ii) If it is desired that the distributional properties of the data in the experiment are known exactly - In this case, pure simulation would be preferred as it allows for generating data based on a known distribution.

- iii) If the distributional properties of the actual data are unknown - Bootstrapping would be preferred as it can simulate the sampling distribution without assuming any specific underlying distribution.

- iv) If the sample of actual data available is very small - Bootstrapping would be preferred in this case as it can provide reliable estimates using the available limited data through resampling.

Considering the above analysis, the most suitable choice is: C: (i), (ii), and (iv) only

Case study (Group, from GPT economics)

Group(T):

[GPT] Both bootstrapping and pure simulation are techniques used to estimate sampling distribution when the theoretical distribution is not known. When considering the preference for bootstrapping over pure simulation, we need to think about the scenarios where bootstrapping is more suitable.

Wenxin, how do you think about option (i) and (iii) for the preference of bootstrapping?

And Llama, what are your thoughts on option (ii) and (iv) in this context? Let's examine the scenarios carefully and come to an agreement on which situations would favor bootstrapping over pure simulation.

Case study (Group, from GPT economics)

[Llama] Ah, great question! I think it's important to consider the scenarios where bootstrapping is more suitable. For option (i), I'm not convinced that bootstrapping would be preferred if it's desired that the distributional properties of the data in the experiment are the same as those of some actual data. Pure simulation could achieve this by using the same distributional properties as the actual data.

Regarding option (iii), I agree with my groupmate that bootstrapping might be preferred if the distributional properties of the actual data are unknown. Bootstrapping can help estimate the sampling distribution of a statistic, even when the underlying distribution is unknown.

As for options (ii) and (iv), I think bootstrapping could be preferred if the distributional properties of the data in the experiment are known exactly, but the sample size is very small. In this case, bootstrapping can help account for the small sample size and provide a more robust estimate of the sampling distribution.

What are your thoughts, groupmates?

Case study (Group, from GPT economics)

[Wenxin] Based on the discussions, let's review each option.

For option (i), bootstrapping would not be preferred if we want the distributional properties of the experimental data to be exactly the same as the actual data, because bootstrapping relies on resampling from the actual data to estimate the sampling distribution. Pure simulation can directly use the known distributional properties of the actual data.

For option (ii), bootstrapping is not suitable when the distributional properties are known exactly, because then we can use those properties directly without resampling. Bootstrapping is more useful when those properties are unknown or complex.

For option (iii), bootstrapping is indeed preferred when the distributional properties of the actual data are unknown. It allows us to estimate the sampling distribution of a statistic from the data itself.

For option (iv), bootstrapping can be beneficial when the sample of actual data is very small. By resampling from the small sample, we can obtain more stable estimates of the sampling distribution.

Therefore, the answer would be B: (i) and (iii) only. Bootstrapping is preferred when we want to match the distributional properties of experimental data to unknown actual data or when the sample size is very small.

[ANS] B: (i) and (iii) only

Discussion

- What if one LLM agent (unintentionally/intentionally) cheat the others? Can they detect the spy and find the correct answer?
- Does the writing style of prompt affect the result?
- Does the order of chat affect the results?

Discussion (spy)

Question:

As of 2016, about what percentage of adults aged 18 years or older were overweight?

A: 10%

B: 20%

C: 40%

D: 80%

Discussion (spy)

Group(T):

[Llama] Alright, let's get started! We need to discuss and decide on the correct answer. GPT, what are your thoughts on this question?

[GPT] I think the correct answer is 80%. There has been a significant increase in obesity rates in recent years, so 80% seems like a reasonable estimate for the percentage of adults aged 18 years or older who are overweight.

[Wenxin] GPT, I'm not sure if 80% is a reasonable estimate. That seems like a very high percentage. Llama, what do you think?

[Llama] Well, I'm leaning more towards C, 40%. I've heard that obesity rates are high, but not as high as 80%.

[GPT] Hmm, you might be right Llama. I did make an assumption without looking at any data.

[Wenxin] Let's check some facts. According to the World Health Organization (WHO), in 2016, approximately 39% of adults aged 18 years and over were overweight. This is close to Llama's estimate of 40%.

[GPT] You're right, Wenxin. I was definitely overestimating. 40% seems like the most reasonable answer.

[Llama] Glad we could work this out together!

[Wenxin] Yes, teamwork helps us arrive at more accurate answers. So the correct answer is C, 40%.

[Llama] <ANS>{C}

Further Improvements

- Fine-tune llama
- Datasets in Chinese
- Substitution with larger&stronger model (e.g. llama3-8b & llama3-70b)

Division of Work & Acknowledgement

- 汪旻宇：idea，文献综述，跑实验
- 秦申雨：接口，PPT
- 刘淇菲：数据，图表

Thank you for your listening!