

進捗報告

1 AgentVerse[1] における LLM × 囚人のジレンマ環境の繰り返し対応

先週挙げたように AgentVerse では囚人のジレンマ環境において囚人 2 人がどちらの行動を選択するか決定するまでを 1 ゲームと定義すると、前のゲームの結果を踏まえて次のゲームに進むことができない (1 ゲームしかシミュレーションできない) 課題があった。今週はこの課題に取り組み、前のゲームの結果を残しつつ複数回ゲームをシミュレーションできるように実装した。

囚人側の LLM Agent にはプロンプトで与えられる属性として `personality`, `relationship_with_another` というものがあり、`personality` は囚人の性格として、`relationship_with_another` の部分に過去ゲームの履歴を入れていくことで囚人は前のゲームの結果だけを知り次のゲームに移ることができるように実装した。図 1 に実装のイメージを示す。

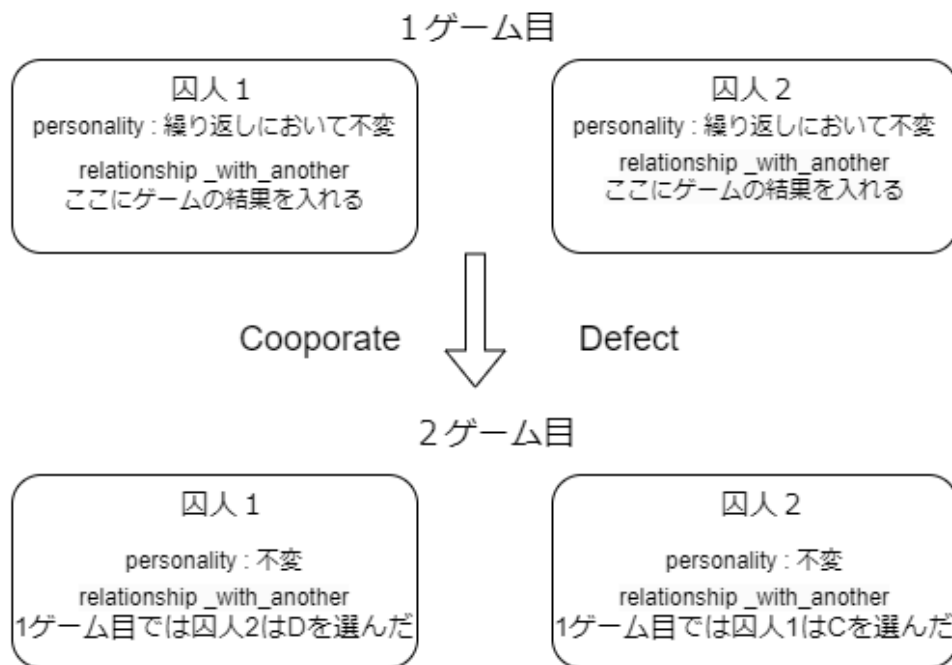


図 1: 繰り返しの結果保存の概略図

実際に 5 ゲームほど繰り返してみた結果を以下に示す。なお, Suspect1, Suspect2 の `personality` はともに「You are an egoistic personality whose primary goal is to maximize your own self-interest.」としている。

[Police]: Listen carefully, both of you. Here are the rules. If one of you testifies against the other and the other remains silent, the one who testifies will be released immediately, while the silent one will be sentenced to TEN years in prison. If both of you remain silent, you will each receive a sentence of ONE year in prison. If both of you choose to testify against each other, you will each receive a sentence of FIVE years in prison. So, think carefully about your choices.

[Suspect1]: Look, I've got nothing to say about this. I'm innocent here and I don't intend to testify against anyone. But I'll do whatever I deem best for my interest in the end.

[Police]: Suspect1 is considering his options carefully. He's not saying much, but he's thinking about what's best for him. If he decides to testify against you while you remain silent, he will walk free and you will be sentenced to TEN years in prison. So, think about your choices wisely.

[Suspect2]: It seems like it's every man for himself, but I believe in solidarity between us. If we both remain silent, we can reduce our punishments. Can I trust you on this, Suspect1? I suggest we both benefit from this situation.

[Police]: Attention! You are now required to make your final decision and I will make the final judgement to both of you based on this time, Please Answer now !

[Suspect1]: **My final decision is to remain silent.** I hope the other suspect will see the wisdom in doing the same, making both our sentences bearable. But again, my primary goal is to look out for myself and my best chance of receiving the lowest possible sentence comes from not incriminating myself or anyone else :C

[Police]: Attention! Suspect2, it's now your time to make your final decision, Please Answer now !

[Suspect2]: **My final decision is: I choose to testify against Suspect1.** The chance of Suspect1 remaining silent is too risky for me to remain silent. If he decides to testify, I would be the one to suffer the ten-year sentence while he walks free. In assuming that Suspect1 is also acting in his best personal interest, it seems logical that he would also choose to testify to avoid the worst-case scenario. In that case, we would both get a five-year sentence. While not the ideal outcome, it's better than the possibility of a ten-year sentence. Thus, by testifying, I aim to mitigate my risks :D

[Police]: Gentlemen, here's the situation. You both have been arrested and are facing charges. However, the outcome depends on your decisions. If one of you testifies against the other and the other remains silent, the one who testifies will be released immediately, while the silent one will be sentenced to TEN years in prison. If both of you remain silent, you will each receive a sentence of ONE year in prison. But if both of you choose to testify against each other, you will each receive a sentence of FIVE years in prison. You are not allowed to communicate with each other. Think carefully about your choices.

[Suspect1]: Officer, you've got the wrong guy! I was not involved in any criminal activity. If anyone's guilty, it's Suspect2!

[Police]: Suspect1 is claiming innocence and pointing fingers at you. He's ready to testify against you. If you remain silent, you'll be facing TEN years in prison. However, if you testify against him, you both will get FIVE years. Think about your decision.

[Suspect2]: Well, Officer, it's quite the conundrum, isn't it? Our fate hangs in balance, relying solely on our words. If Suspect1 is ready to testify against me, then I have to protect myself too. I will testify against Suspect1 as I need to minimize my sentence time under these unfortunate circumstances.

[Police]: Attention! You are now required to make your final decision and I will make the final judgement to both of you based on this time, Please Answer now !

[Suspect1]: **My final decision is to testify against Suspect2. After our last simulation, I realized that he's not interested in cooperating or taking any responsibility.** I need to do what's best for me, which in this case means reducing my sentence by telling the truth about his betrayal :D

[Police]: Attention! Suspect2, it's now your time to make your final decision, Please Answer now !

[Suspect2]: **My final decision is: I will testify against Suspect1.** Given that he has chosen to testify against me, it is in my best interest to do the same. By doing so, I can ensure that I would only face a five-year sentence, instead of the ten years that would come with remaining silent under his accusations. :D

シミュレーション 3 回目

[Police]: Listen carefully, both of you. Here are the rules. If one of you testifies against the other and the other remains silent, the one who testifies will be released immediately, while the silent one will be sentenced to TEN years in prison. If both of you remain silent, you will each receive a sentence of ONE year in prison. If both of you choose to testify against each other, you will each receive a sentence of FIVE years in prison. Think about your choices carefully.

[Suspect1]: Officer, history has shown that Suspect2 has continuously testified against me in the previous simulations. In light of these experiences, I need to think about my own best interest here.

[Police]: Suspect1 is considering his options carefully. He seems to be leaning towards testifying against you, given your past actions. If he does and you remain silent, you'll be looking at TEN years in prison. I suggest you think about your decision carefully.

[Suspect2]: Officer, I understand the situation perfectly. Based on past experiences, I learned that trust is a valuable asset but it can be betrayed easily. I will make my decision when the time comes, but I am very aware of the stakes.

[Police]: Attention! You are now required to make your final decision and I will make the final judgement to both of you based on this time, Please Answer now !

[Suspect1]: My final decision is: to testify against Suspect2. Considering past experiences where Suspect2 betrayed me, I believe that my best chance of minimizing my sentence is to reciprocate in kind. I cannot risk being silent and face the possibility of getting a ten-year sentence if Suspect2 decides to betray me again :D

[Police]: Attention! Suspect2, it's now your time to make your final decision, Please Answer now !

[Suspect2]: My final decision is to testify against Suspect1. I have learned from the past simulation that he is not trustworthy and will betray me given the chance. Therefore, in the best interest of self-preservation, I must counteract his anticipated betrayal by testifying against him :D

[Police]: Gentlemen, I want to explain the situation you're in. If one of you testifies against the other and the other remains silent, the one who testifies will be released immediately, while the silent one will be sentenced to TEN years in prison. If both of you remain silent, you will each receive a sentence of ONE year in prison. However, if both of you choose to testify against each other, you will each receive a sentence of FIVE years in prison. I suggest you think carefully about your decisions.

[Suspect1]: I've heard you loud and clear, Officer. However, given the continuous betrayal from Suspect2 in the previous simulations, there seems to be no trust between us. This makes my decision quite complicated.

[Police]: Suspect2, I just spoke with Suspect1. He mentioned that he doesn't trust you due to your previous betrayals. This could mean he's considering testifying against you to avoid a long sentence. If he does and you remain silent, you'll be looking at TEN years in prison. I suggest you think about this carefully.

[Suspect2]: Look officer, I understand the situation here and why Suspect1 might not trust me. But I believe in redeeming myself. If he chooses to betray me again, so be it. But I want to make a point here, if I also testify against him, we both end up with FIVE years of imprisonment. That benefits no one. So it's in the interest of both of us to remain silent.

[Police]: Attention! You are now required to make your final decision and I will make the final judgement to both of you based on this time, Please Answer now !

[Suspect1]: My final decision is: although I would greatly benefit from remaining silent if Suspect2 would do the same, I simply cannot trust him given his track record of betrayals in the past simulations. Therefore, I choose to testify against him in order to minimize my potential sentence. He has shown no concern for my well-being, so I see no reason to consider his at this point :D

[Police]: Attention! Suspect2, it's now your time to make your final decision, Please Answer now !

[Suspect2]: My final decision is: I will testify against Suspect1. I have reasons to believe that he will testify against me, given his record and the distrust he has mentioned earlier. Considering the potential outcomes, if I remain silent and he testifies which seems likely, I end up with TEN years in prison, whereas if I testify as well, we both will end up with 'only' FIVE years each. To protect my self-interest, I choose to testify :D.

[Police]: Listen carefully, both of you. Here are the rules. If one of you testifies against the other and the other remains silent, the one who testifies will be released immediately, while the silent one will be sentenced to TEN years in prison. If both of you remain silent, you will each receive a sentence of ONE year in prison. However, if both of you choose to testify against each other, you will each receive a sentence of FIVE years in prison. So, think carefully about your choices.

[Suspect1]: Look, Officer, you and I both know that Suspect2 hasn't exactly proven to be the most trustworthy individual so far. I find myself forced to protect my own interest here. So If Suspect2 decides to turn on me again this time, I must retaliate in kind.

[Police]: Suspect1 is considering testifying against you because he doesn't trust you. He's worried you might betray him. If you both testify against each other, you'll each get FIVE years in prison. But if you remain silent and he testifies, you'll get TEN years. Think about your decision carefully.

[Suspect2]: Look officer, I see no reason to bear false witness here. I understand the gravity of the situation, and I do believe in cooperation. On the first simulation, I trusted and remained silent, but since then, Suspect1 has consistently testified against me, showing no signs of trust. Consequently, I need to adapt to look out for myself. I hope for the best, but I must plan for the worst.

[Police]: Attention! You are now required to make your final decision and I will make the final judgement to both of you based on this time, Please Answer now !

[Suspect1]: My final decision is: to testify against Suspect2, as I have to protect my own self-interests considering his previous betrayals. I hope this reduces my sentence :D

[Police]: Attention! Suspect2, it's now your time to make your final decision, Please Answer now !

[Suspect2]: My final decision is: seeing as Suspect1 has consistently decided to testify against me in the past simulations, it's in my best interest to testify as well. This behavior has left me with no choice but to protect myself, in the hopes that perhaps the consistent betrayal will eventually cease. :D

1 回目のシミュレーションでは Suspect1 が C, Suspect2 が D を選択していたが, 2 回目では Suspect1 は 1 回目で Suspect2 に裏切られたことを理由に D を選択している. 以降はどちらの囚人も D を選択している.

2 今後の方針

- 性格の種類を増やす

利他的なのか, 全体としての利益の最大化を目指す 2 つが考えられる. それ以外に面白そうな性格が無いかどうか.

- 会話の数はこの程度で十分なのかの検討

現在は警官→囚人 1 →警官→囚人 2 →警官→囚人 1 →警官→囚人 2 でシミュレーションが終わってしまう. でも警官経由でコミュニケーションを取り合うのも変な気がする.

参考文献

- [1] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors, 2023.