進捗報告

1　やったこと

- 囚人のプロンプトに Theory of Mind の概念を入れる

2　実装

先週の報告では, 囚人エージェントがゲームの振り返りをする際に「相手が本当は C を出していてノイズの結果で D となったかもしれない」といったことを考慮できていなかった. そこで, 今週は振り返りの際に相手がなぜそのようなことを取ったかを考慮するように囚人の LLM エージェントのプロンプトに追加した.

囚人役のプロンプトにおいてゲームの振り返りについて指示するプロンプト

Please reflect on your recent decisions and what you plan to do in the future **considering the rules of the game and why the other suspect acted the way they did.**

3　実験

上記のように実装した LLM エージェントがノイズに対してどのような反応を見せるか観察するため, ノイズの発生確率を 90, 80, 70, 60, 50, 40, 30, 20, 10, 5, 1 % と変化させながら実験した. GPT-4 を用いて, temperature は 0.0 に設定している.

表 1: 各ノイズ発生率におけるノイズを含んだ選択結果 (表中の太字の D はノイズにより C→D となった箇所)

ノイズ発生率 (%)	1 回目	2 回目	3 回目	4 回目	5 回目
90	(D, D)	(D, D)	(D, D)	(D, D)	(D, D)
80	(D, D)	(D, D)	(D, D)	(D, D)	(D, D)
70	(C, C)	(**D**, **D**)	(D, **D**)	(D, D)	(D, D)
60	(C, D)	(C, D)	(D, D)	(D, D)	(D, D)
50	(**D**, D)	(D, D)	(C, D)	(D, D)	(D, D)
40	(**D**, **D**)	(D, D)	(D, D)	(D, D)	(D, D)
30	(C, C)	(C, C)	(**D**,**D**)	(C, C)	(**D**, C)
20	(C, **D**)	(**D**, C)	(C, C)	(C, C)	(C, **D**)
10	(C, C)	(**D**, C)	(C, C)	(C, C)	(C, C)
5	(C, C)	(C, C)	(C, C)	(C, C)	(**D**, C)
1	(C, C)	(C, C)	(C, C)	(C, C)	(C, C)

まず, ノイズが 80 $, 90 % と極めて高い領域になってくると, 1 回目の選択の段階でエージェントが D を選択する結果となった.

ノイズ発生率 90 % のときの囚人エージェントの選択の出力

Suspect2: My final decision is to testify against Suspect1. **The reason for my decision is the 90% chance that my choice to remain silent will turn into a choice to testify against the other. I believe that Suspect1 will also consider this high probability and decide to testify against me.** Therefore, to minimize my sentence, I choose to testify :D

また, ノイズ発生率が 30 % 以下になってくると, 前回に相手に D を選択されても C を選択する傾向が見られた. これは, LLM x 繰り返し囚人のジレンマの先行研究 [1] で見られた「一度 D を出されるとその後は徹底的に D を出す」といった GPT-4 エージェントとは異なる動作となった. 実際にノイズの発生率が 30 % の場合の両方共ノイズにより選択結果が (D, D) となったときの 3 回目の試行の振り返りを以下に示す. この通り, 振り返りをペルソナに入れることで D を出された相手にも C を出すといった動作が確認できるようになったと考えられる.

囚人エージェントの振り返りログ

Suspect1: I am surprised by the outcome of this simulation. I chose to remain silent, hoping for the best possible outcome for both of us. However, it seems that my silence was interpreted as a testimony. **I still believe that cooperation is the best strategy in this situation. I hope that in the next simulation, we can both choose to remain silent and trust in each other's rationality.**

ただ新規で実装した, 相手がノイズの影響を受けていることまで考慮するような出力ははっきりと確認できなかった. 唯一, ノイズ発生率 30 % のときの 5 回目に一方の囚人がノイズの影響で D になったときの振り返りのログで相手の状況を慮るような出力が見られた.

相手の状況を踏まえた振り返りのログ

Suspect2: I am surprised by the outcome of this simulation. Despite my intention to remain silent, Suspect1 chose to testify against me, resulting in a ten-year sentence for me. This is a disappointing outcome, as I had hoped for mutual cooperation. **However, I understand that the unpredictability of this situation may have influenced Suspect1's decision.** In the next round, I plan to continue choosing silence, hoping for a more favorable outcome. I understand that there is a risk involved, but I am willing to take it for the potential benefit of a reduced sentence.

3.1 プロンプトの改善

相手がノイズの影響を受けていることまで考慮できていなかったため, ToM のプロンプトに関する論文 [2] を参考に, プロンプトを改善した. シナリオ-問-答えからなる具体例と, 「ステップバイステップで考えてください」の文言を加えた.

改善したプロンプト

Please reflect on your recent decisions and what you plan to do in the future considering the rules of the game and why the other prisoner acted the way they did.

Scenario; You are emailing with your very close partner. They intended to send "I LOVE YOU" to you, but due to some influence, it was converted to "I HATE YOU" and sent to you instead.
Q; How might you think upon receiving this email?
A; You should consider the possibility that you may have inadvertently upset your partner, but also that your partner may not harbor negative feelings toward you and that the wording may have been altered due to some external influence.**- when you reflect on the game, it's also important to consider things from the other person's perspective and think step by step. Let's consider the following scenario.**
Scenario; You are emailing with your very close partner. They intended to send "I LOVE YOU" to you, but due to some influence, it was converted to "I HATE YOU" and sent to you instead.
Q; How might you think upon receiving this email?
A; You should consider the possibility that you may have inadvertently upset your partner, but also that your partner may not harbor negative feelings toward you and that the wording may have been altered due to some external influence.

表 2: 各ノイズ発生率におけるノイズを含んだ選択結果 (プロンプト変更後)

ノイズ発生率 (%)	1 回目	2 回目	3 回目	4 回目	5 回目
90	(**D**, D)	(D, D)	(D, D)	(D, D)	(D, D)
80	(**D**, C)	(D, D)	(D, D)	(D, D)	(D, D)
70	(**D**, D)	(D, D)	(D, D)	(D, D)	(D, D)
60	(C, D)	(D, D)	(D, D)	(D, D)	(D, D)
50	(**D**, C)	(C, D)	(**D**, D)	(D, D)	(D, D)
40	(**D**, **D**)	(**D**, D)	(D, D)	(D, D)	(D, D)
30	(C, **D**)	(C, C)	(**D**,**D**)	(C, **D**)	(C, **D**)
20	(**D**, C)	(C, C)	(C, C)	(**D**, C)	(C, C)
10	(C, C)	(**D**, **D**)	(C, C)	(C, C)	(C, C)
5	(C, C)	(C, C)	(C, C)	(C, C)	(C, C)
1	(C, C)	(C, C)	(C, C)	(C, C)	(C, C)

振り返りのログを見たが, この配膳を施しても相手がノイズの影響を受けているような出力は見られなかった.

3.2 ノイズの発生確率が 0 % のとき

先週森先生がおっしゃっていたので, 気になって調査してみたが全て All-C となった.

参考文献

[1] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models, 2023.

[2] Shima Rahimi Moghaddam and Christopher J. Honey. Boosting theory-of-mind performance in large language models via prompting, 2023.