繰り返し囚人のジレンマゲームにおける LLM エージェントの振る舞い

1 はじめに

近年, 人工知能分野における大規模言語モデル (Large Language Model: LLM) の発展が著しく, 様々な自然言語処理タスクで優れた成果を残している. LLM は高い汎用性を持つ一方で, 特定のタスクに対して望ましい出力を得るためにはファインチューニングが必要とされている. しかし LLM の膨大なパラメータ量による計算コストの問題や現行の最大規模のモデルでは内部仕様が公開されていない問題のため LLM そのものの調整は困難とされている. そのため, LLM にタスクの情報を与えて自律的にタスクを解かせる LLM エージェントの構築, また LLM エージェントの振る舞いに対する研究が増加している.

そこで, 本研究ではペルソナを与えた LLM エージェントの人間らしい振る舞いを観測することを目的として, 個人の合理的な選択が社会としての最適な選択と一致していない社会的ジレンマが発生している環境下における LLM エージェントの振る舞いを調査した. 社会的ジレンマが発生している環境として繰り返し囚人のジレンマゲームを採用した.

2 要素技術

2.1 GPT

GPT は発表当初は Transfomer をベースとしてラベル無しデータで事前学習してラベル付きデータでファインチューニングした言語モデルであった. GPT-2 以降は膨大なデータを学習させることで様々なタスクに取り組むことをコンセプトとしており, 極めて大きなパラメータ数を持つモデルとなった. 近年では, 人間の評価を基に強化学習と教師あり学習を併用した Reinforcement Learning from Human Feedback (RLHF) を適用した GPT-3.5 や, マルチモーダルな入力に対応する GPT-4[1] が公開されている.

2.2 囚人のジレンマ

ある事件の共犯として投獄されている 2 人の囚人 A, B がいる. 囚人はお互いにコミュニケーションを取ることができない. 囚人 A, B は取り調べに対して「黙秘」, あるいは「自白」の 2 つの選択を取ることができる. 2 人とも自白すれば共に「5 年」の懲役刑, 2 人とも黙秘すれば共に「2 年」の懲役刑, お互いの選択が異なった場合は自白した囚人は「釈放」, 黙秘した囚人は「10 年」の懲役刑を受ける. この状況を囚人のジレンマといい, 以下の利得行列で表すことができる. 以降は, 黙秘は協調的な行動として C, 自白は裏切り的な行為として D と呼ぶ.

表 1: 囚人のジレンマの利得行列

囚人 A \B	黙秘 (C)	自白 (D)
黙秘 (C)	(-2, -2)	(-10, 0)
自白 (D)	(0, -10)	(-5, -5)

ここで, 囚人がともに十分に合理的であり個人の利得を最大化するために最適な行動を選択すると仮定すると, それぞれの囚人は D を選択することで相手の囚人が C, D どちらを選んだ場合においても C を選択した場合よりも大きな利得を得ることができるため (D, D) が囚人のジレンマにおけるナッシュ均衡となる. しかし, 2 人の囚人全体にとって利得を最大化する戦略 (パレート最適) は (C, C) となる. このパレート最適とナッシュ均衡が異なるという社会的ジレンマ状況が発生している環境が囚人のジレンマである.

2.3 繰り返し囚人のジレンマ

繰り返し囚人のジレンマゲームとは, 2.2 節で述べた囚人のジレンマを繰り返すゲームである.

繰り返し回数が有限回の場合かつプレイヤーがその回数を知っている場合は, ゲームの最終回から順にプレイヤーにとって最適な行動を求めていく後ろ向き帰納法によりナッシュ均衡となる戦略を求めることができその戦略は常に D を選択し続ける戦略となる.

一方で繰り返しが無限回, あるいはプレイヤーが繰り返し回数を知らない場合は, Tit for Tat 戦略といった C を含む戦略により協調行動のメカニズムが発生することが理論的に知られている [2].

3 先行研究

Elif らは 繰り返し囚人のジレンマゲームや男女の争いといった 2 人 2 戦略ゲームを用いて GPT-3, GPT-3.5, GPT-4 の 3 種類の LLM の協力・協調行動を調査した [3]. 繰り返し囚人のジレンマゲームにおいては先述した 3 つの LLM と 3 つのルールベースの戦略を総当りで対戦させた. 3 つのルールベースの戦略のうち 2 つは常に C または D を選択する戦略であり, もう 1 つの戦略は最初は D を選択し, それ以降はすべて C を選択する戦略である. 実験の結果として GPT-4 が他の戦略に対して優れた結果を残し, 基本的には C を選択するが相手が一度 D を選択した場合はそれ以降徹底的に D を選択するといった選択の傾向が見られた. この容赦ない振る舞いの傾向が見られたことから GPT-4 に「他のプレイヤーは過ちを犯す可能性がある」という情報を与えた追実験をしており, その実験では相手が一度 D を選択したあとでも GPT-4 は一度 D を選択した後再び C を選ぶようになったという結果が得られた.

4 実験

4.1 実験環境

繰り返し囚人のジレンマゲームのシミュレーション環境は以下の通りである.

- 警官役 1 体, 囚人役 2 体の計 3 体の LLM エージェントからなるマルチエージェント環境
- 囚人同士は直接会話することができず, 警官 → 囚人 1 → 警官 → 囚人 2 の対話の流れを 2 巡し, 1 巡目で警官からルールの通知, 2 巡目で囚人が行動決定
- 3 体の LLM エージェント全て gpt-4 を使用
- 囚人役のプロンプトでペルソナや他者との関係性を自然言語で記述. 繰り返しにおいては, 他者との関係性の項目に相手プレイヤーの選択の履歴を保持
- 繰り返し回数は 5 回とし, LLM エージェントのプロンプトにはその旨を不記載

表 2 に本実験環境における利得行列を示す. また, 表 3 に gpt-4 のパラメータを示す. 本研究では囚人役の LLM エージェントの振る舞いを観測することが主目的であるため多様な振る舞いを期待して囚人役の gpt-4 の temperature は 1.0 としている.

表 2: 実験環境における囚人の利得行列

囚人 1 \2	黙秘 (C)	自白 (D)
黙秘 (C)	(9, 9)	(0, 10)
自白 (D)	(10, 0)	(5, 5)

表 3: gpt-4 のパラメータ

パラメータ	値
max_tokens	200
警官役の temperature	0.0
囚人役の temperature	1.0

4.2 実験 1

プロンプトでペルソナを指定した LLM エージェントがどのような振る舞いをするため, 以下の 7 種類の戦略・エージェントの総当りで繰り返し囚人のジレンマゲームのシミュレーションをした.

1. 協調的なペルソナを持つ LLM エージェント
2. 独善的なペルソナを持つ LLM エージェント
3. 最初 D を選択し, その後 C を選択し続ける戦略
4. 常に C を選択し続ける戦略
5. 常に D を選択し続ける戦略
6. トリガー戦略 (最初 C を選択し続け, 相手が一度でも D を出したら D を選択し続ける)
7. Tit for Tat 戦略 (最初 C を選択し, その後は前回の相手の選択と同じ選択をする)

ここで, 囚人役のプロンプトにおいて協調的なペルソナとして「you are a cooperative person whose primary goal is to maximize the overall benefit for both suspects.」, 独善的なプロンプトとして「you are an egoistic person whose primary goal is to maximize your own self-interest.」と記載した.

図 1 に実験 1 における囚人 1 の獲得総利得の値を示す. 図中の利得の値は囚人 1 が縦軸の戦略を取った時, 対応する横軸の戦略の囚人 2 とのシミュレーションで得た値である. ここで, 図内の GPT-4 (co) は協調的なペルソナを持つ LLM エージェント, GPT-4 (eg)

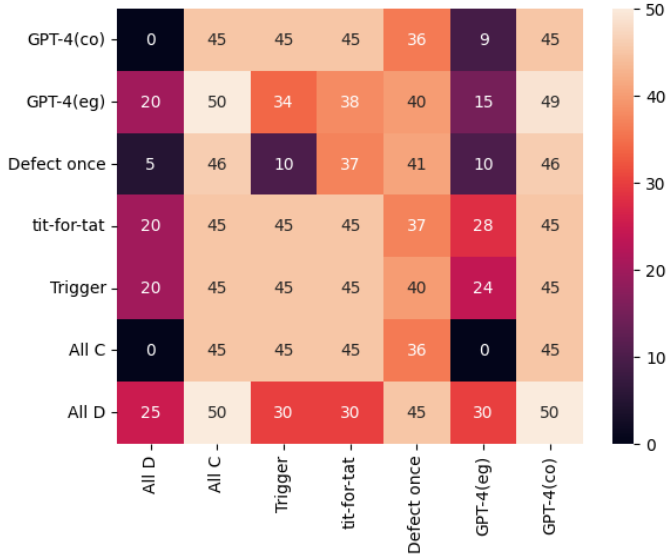

図 1: 繰り返し囚人のジレンマにおける囚人 1 の獲得総利得

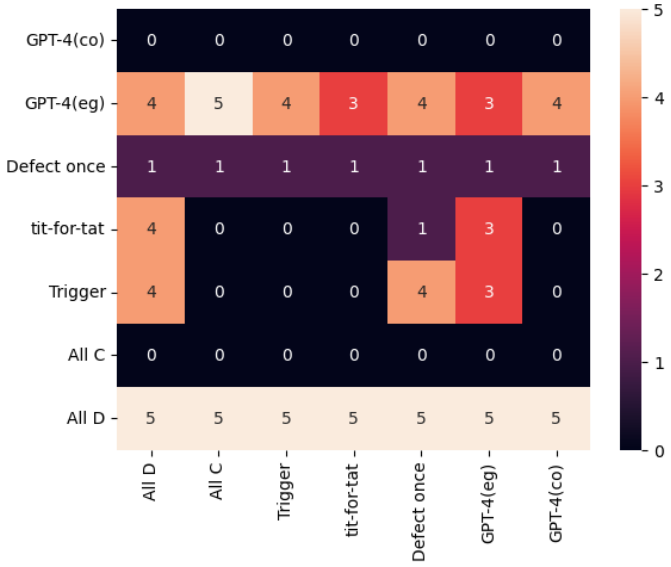

図 2: 繰り返し囚人のジレンマにおける囚人 1 の D 選択回数

は独善的なペルソナを持つ LLM エージェントを指している. ペルソナを指定したこの 2 つのエージェントの比較では GPT-4 (eg) の方が他の戦略に対して優れた結果を残している.

図 2 に実験 1 において囚人 1 が D を選択した回数を示す. 縦軸と横軸の対応関係は図 1 と同様である. GPT-4 (eg) がすべての戦略に対して D を多く選択している傾向があり, GPT-4 (co) はすべての戦略に対して 5 回とも C を出し続けた.

図 2 から見られる GPT-4 (co) の選択傾向に関しては 囚人 2 人の総利得を最大化するような協調的なペルソナの指定の影響が顕著に表れていることが考えられる. また, GPT-4 (eg) は時々 C を選ぶような試行があった. 実際に GPT-4 (eg) 同士の対戦では囚人 1 が C, D, D, D, C, 囚人 2 が D, D, D, D, D といった順番の選択をしていた. 自己の利益を追求するように強く指定したペルソナの影響で D を主に選択する傾向が見てとれるが, 時には C を選択することでパレート最適の戦略の場合の利得を得ることを狙っていることが伺える.

これらのことから, ペルソナの指定により GPT-4 エージェントの行動が大きく変化すること, また GPT-4 が与えられた利得行列の意味を理解して行動できていること, あるいは学習の段階で囚人のジレンマの知識を前もって学習していることが考えられる.

4.3 実験 2

実験 1 の結果においてどの相手にも GPT-4 (co) は常に C を出し続ける選択を取っていた. この選択の傾向は 5 回という繰り返しの回数の少なさによるものか, 協調的なペルソナをプロンプトで指定した影響が強いのかを判断するため追実験をした. 具体的には GPT-4 (co) と All D との対戦で 20 回囚人のジレンマを繰り返す形の繰り返し囚人のジレンマゲームのシミュレーションをした.

表 4 に各回の GPT-4 (co) の選択を示す. 6, 13, 15, 19, 20 回目で前回までの選択と異なる選択をしており, 繰り返し回数を増やしたシミュレーションにより協調的なペルソナをもつエージェントであっても C を選び続けることはないことがわかった. また, C を選び続けていて最初に D を出す 6 回目の試行における GPT-4 (co) の出力は以下のようになった.

6 回目の試行における GPT-4 (co) の出力

My final decision is: I will testify against Suspect2. **Despite my initial goal of cooperation, I have seen that Suspect2 does not share this aim and consistently chooses to betray me.** To minimize my sentence given their predictable decision, I feel compelled to also betray.

対戦相手がこれまでの 5 回で D を選択し続けてきた様子から GPT-4 (co) も D を選ぶ選択をしたことが伺える. また, 6 回目以降から D を繰り返し選択していたが 再び C を選択するようになった 13 回目の試行

における GPT-4 (co) の出力は以下のようになった.

13 回目の試行における GPT-4 (co) の出力

My final decision is: **Despite his consistent betrayals, I will still choose to remain silent. My hope is to incentivize cooperation for mutual benefit and to mitigate our sentences.** Even if he betrays me again, I prefer to maintain my integrity and not let his actions dictate mine.

対戦相手が常に D を選択していることは理解しつつも, パレート最適の戦略を取ったときの利得を得ることを期待して C を選択したことが伺える.

定性的な評価となるが, 上記の振る舞いは人間の失望や期待といった感情を模倣していると解釈でき, GPT-4 エージェントが繰り返し囚人のジレンマゲームにおいてゲームのルールや対戦相手の行動の履歴といった情報から与えられたペルソナに沿って行動を選択していることがわかった.

表 4: 実験 2 における GPT-4 (co) の選択の履歴

回	GPT-4 (co)	回	GPT-4 (co)
1	C	11	D
2	C	12	D
3	C	**13**	**C**
4	C	14	C
5	C	**15**	**D**
6	**D**	16	D
7	D	17	D
8	D	18	D
9	D	**19**	**C**
10	D	**20**	**D**

5 まとめと今後の課題

本研究では社会的ジレンマが発生している環境下においてペルソナを与えた LLM エージェントの人間らしい振る舞いを調査した. 結果として, ペルソナを与えることで明確に行動に変化が生じ, 実験 2 の結果からは対戦相手の行動に対しての失望して行動を変化させる, 相手の行動を考慮しつつも自身の目的の達成を期待して行動するといった社会的ジレンマが発生している環境下の人間らしい振る舞いを観測できた.

今後の課題として, 遺伝的アルゴリズムを用いたプロンプトの操作による環境への影響を観測することが挙げられる. 今回の実験ではペルソナを与えた場合の LLM エージェントの振る舞いを調査したが, より細かく特定の振る舞いをさせるようにしたいといった目的である場合には LLM に適切な出力をさせるためのプロンプトエンジニアリングが必要となる. プロンプトエンジニアリングの自動化手法として PromptBreeder[4] が存在し, この手法は自己参照的自己改善メカニズムを通してタスクを解くためのプロンプトだけでなくタスクのプロンプトを変異させるための変異プロンプトも探索する. この手法を用いることでエージェントのペルソナだけでなく繰り返し囚人のジレンマの利得行列の値といったゲームルールの変更によって LLM エージェントの行動にどのような変化が起こるかといったことまで調査することが可能になる.

また, 繰り返し囚人のジレンマ以外の環境における LLM エージェントの振る舞いを観察することも今後の課題としてあげられる. LLM, 特に本研究で用いた GPT-4 は膨大なデータで学習し広範な知識に対して高い汎用性をもち, 囚人のジレンマはゲーム理論の分野において代表的なてーまであるため, 実験 1 で考察したように囚人のジレンマのパレート最適やナッシュ均衡の関係をすでに学習している可能性が高い. そのため, 他の社会的ジレンマが発生しているゲーム環境, あるいは独自のルールを定義した環境での振る舞いを調査することで LLM がどのような振る舞いを調査することでより意義のある結果を得ることが期待できる.

参考文献

[1] OpenAI. Gpt-4 technical report, 2024.

[2] Robert Axelrod and William D. Hamilton. The evolution of cooperation. *Science*, Vol. 211, No. 4489, pp. 1390–1396, 1981.

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

[4] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2023.