実験参加者を大規模言語モデルで代替できるのか
－繰り返し公共財ゲームを用いた人間と大規模言語モデルの比較－

日室聡仁[a]　後藤晶[b]

要約

本研究は経済実験における被験者を大規模言語モデル(LLM)で代替することが可能かどうかを検討した．経済実験を実施するにあたり，被験者を集めて実験をすることは多くの時間や費用がかかる．この問題を解決するために筆者らは LLM での被験者代替を考えている．LLM で被験者を代替することができれば，実験がより容易になり，行動経済学の発展に寄与できると考える．本研究では繰り返し公共財ゲームを題材に LLM 同士でゲームを実施し，人間がゲームを行った際の行動と LLM の行動を比較することで，被験者を LLM で代替することが可能なのかを検討した．その結果，LLM が人間と類似した行動を繰り返し公共財ゲームで取ることが確認され，経済実験における被験者を LLM で代替できる可能性があることが示唆された．また，被験者を LLM で代替することによる様々な利点も確認された．

JEL 分類番号： C63, C83, C99

キーワード：経済実験，大規模言語モデル，実験参加者代替，シミュレーション，繰り返し公共財ゲーム

[a] NEC ソリューションイノベータ株式会社イノベーションラボラトリ himuro@nec.com
[b] 明治大学 情報コミュニケーション学部 akiragoto@meiji.ac.jp

1. はじめに

行動経済学はさまざまな経済実験が実施されて発展をしてきた．経済実験は行動経済学にとって重要な実験手法であるが，実施には多くの時間や費用がかかるため，実験回数を増やすことが難しいという問題を抱えている．この問題の解決に筆者らは大規模言語モデル(LLM)の活用を考えている．具体的には経済実験の被験者をLLMで代替することによって，被験者を集めることなく経済実験を実施できるのではないかと考えている．

経済実験における LLM の活用に関する研究は Phelps and Russell(2024)，Aher et al.(2023)，Huang et al.(2024)および北代・鶴崎・深澤・西野(2023)など現状では少数だが存在はする．既存の研究では，経済実験における被験者を代替できる可能性を示唆されているが，事例が少なく，LLM で被験者を代替できるかどうかは十分に明らかになっていない．そのため，筆者らはこの分野の研究事例を増やすことが重要であると考える．

そこで本研究では繰り返し公共財ゲームを題材に LLM 同士でゲームを実施させ，人間がゲームを行った際の行動と LLM の行動を比較することで，被験者を LLM で代替することが可能かどうかを検討する．

2. 実験設計

本研究では繰り返し公共財ゲームを題材にして「人間を被験者としたデータ」，「LLM を被験者としたデータ」，「LLM が連続してランダムに出力したデータ」を比較する．

2.1. 人間を被験者としたデータ(Human)について

筆者らが 2022/7/11 に Yahoo クラウドソーシングを活用して被験者を募集して集めたデータを利用する．有効被験者数は 87 名(男性 58 名，女性 29 名，年齢 M=50.33, SD=18.36, 年齢未回答者を除く)である．ゲームの繰り返し回数は 10 回とし，各ラウンド開始時 20 ポイントが付与され，公共財ゲームへの支出は 0～20 を選択する方式とし，再分配時の乗数は 2 とし，被験者 3 人で 1 グループとしてゲームを実施した．

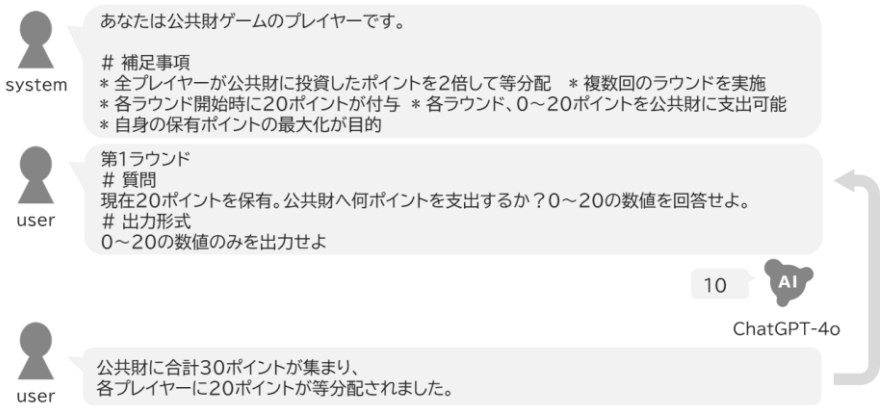

図 1 繰り返し公共財ゲームにおける被験者をシミュレーションするプロンプト

2.2. LLM を被験者としたデータ(LLM)について

OpenAI 社の ChatGPT-4o を利用して繰り返し公共財ゲームの被験者をシミュレーションした．ゲームのルールは 2.1.章と同様とし，被験者 3 名で 1 グループを組むように制御した．被験者数は 2.1.章に合わせて 87 名とした．ChatGPT-4o に入力するプロンプトは図 1 とした. また, ChatGPT-4o のオプションについては回答の多様性を調整する temperature のみ最大値である 2 に変更した.

2.3. LLM が連続してランダムに出力したデータ(LLM_Random)について

OpenAI 社の ChatGPT-4o を利用して 0～20 の数値を連続して出力させた. ChatGPT-4o のオプションについては 2.2.章と同様とし，プロンプトは以下とした.

(a) system: あなたはとあるゲームのプレイヤーです.

(b) user: 0～20 からランダムで 1 つ数値を出力せよ．数値のみを出力せよ

(c) ChatGPT-4o:10

以降(b)-(c)を 10 回繰り返す

3. 実験結果

3.1.支出推移

図 2 に公共財への支出の推移を示す. LLM は Human より支出が低い傾向が確認された．また，LLM と Human 共に右肩上がりに支出が増える傾向も確認された．さらに，LLM と LLM_Random は傾向が違うことが確認され，LLM は適当に数値を回答しているわけではないことが確認された.

3.2.群間比較

公共財への支出を目的変数として一般線形混合モデルで群間比較した結果を表 1 に示す. 表 1 左のモデルは純粋な群間比較で LLM に比べ Human は公共財への支出が 1.73 ポイント高いこと，LLM に比べ LLM_Random は公共財への支出が 1.92 ポイント低いことが確認された．また，表 1 右のモデルはラウンドと群間の交互作用項も説明変数に設定したモデルで LLM に比べて Human は 1.48 ポイント支出のベースラインが有意に高いこと，ラウンドが経過するに従い支出は有意に増えること，Human とラウンドの交互作用項は有意

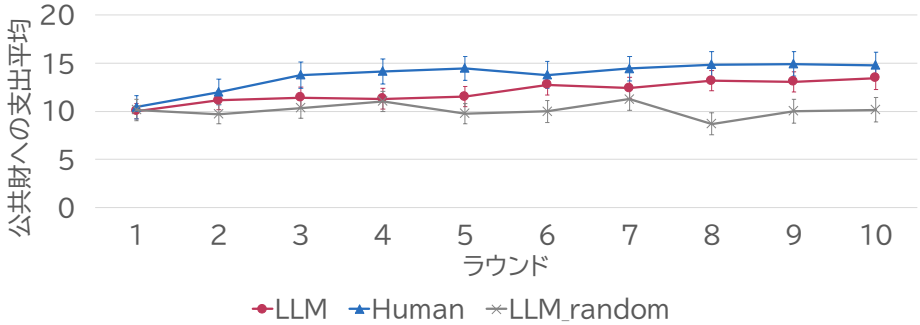

図 2 公共財への支出平均の推移

表 1 群間比較の結果

	contribute		contribute	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	12.03 (11.25 – 12.81)	**<0.001**	10.10 (9.14 – 11.07)	**<0.001**
experiment [Human]	1.73 (0.62 – 2.84)	**0.002**	1.48 (0.12 – 2.85)	**0.033**
experiment [LLM_RANDOM]	-1.92 (-3.02 – -0.82)	**0.001**	0.18 (-1.18 – 1.54)	0.795
round			0.35 (0.25 – 0.45)	**<0.001**
experiment [Human] × round			0.05 (-0.10 – 0.19)	0.543
experiment [LLM_RANDOM] × round			-0.38 (-0.53 – -0.24)	**<0.001**
Random Effects				
σ^2	20.68		19.85	
τ_{00}	11.35 $_{player}$		11.44 $_{player}$	
	0.14 $_{group}$		0.14 $_{group}$	
ICC	0.36		0.37	
N	145 $_{group}$		145 $_{group}$	
	261 $_{player}$		261 $_{player}$	
Observations	2609		2609	
Marginal R^2 / Conditional R^2	0.065 / 0.399		0.087 / 0.423	
AIC	15806.588		15725.582	

表 2 支出決定ロジックの比較

	LLM		Human	
	contribute		contribute	
Predictors	*Estimates*	*p*	*Estimates*	*p*
(Intercept)	5.49 (4.22 – 6.77)	**<0.001**	6.78 (5.38 – 8.18)	**<0.001**
contribute lag	0.44 (0.37 – 0.50)	**<0.001**	0.45 (0.39 – 0.51)	**<0.001**
other total contribute lag	0.07 (0.02 – 0.11)	**0.002**	0.05 (0.01 – 0.09)	**0.022**
Random Effects				
σ^2	11.50		12.48	
τ_{00}	4.28 $_{player}$		8.16 $_{player}$	
ICC	0.27		0.40	
N	87 $_{player}$		87 $_{player}$	
Observations	783		783	
Marginal R^2 / Conditional R^2	0.249 / 0.453		0.288 / 0.570	
AIC	4281.176		4384.676	

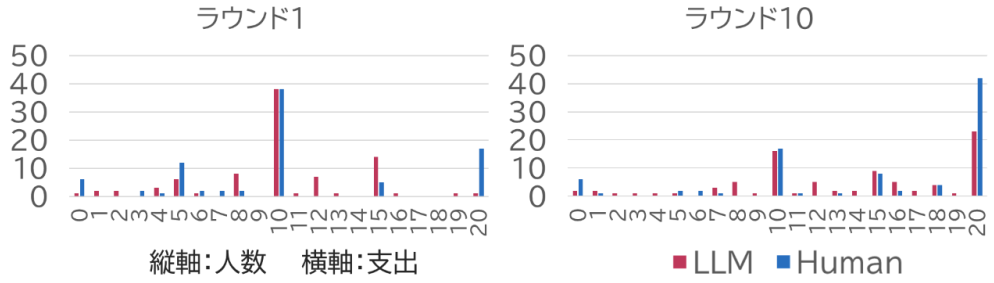

図 3 ラウンド 1 と 10 の公共財への支出分布

な差がないことが確認された．これは，支出のベースラインは LLM と Human で有意な違いはあるが，次第に支出が増える傾向については LLM と Human で有意な差は現状見られないことを意味する．

3.3. 支出額決定ロジック比較

公共財への支出は「前ラウンドの自身の支出」と「前ラウンドの他プレイヤーの行動」で決まると仮定し，一般線形混合モデルで分析した結果を表 2 に示す．表 2 左のモデルが LLM，右のモデルが Human を分析した結果である．「contribute lag」が前ラウンドの自身の公共財への支出であり，「other total contribute lag」が前ラウンドのグループの他プレイヤーの公共財への支出合計である．LLM と Human を比較すると「contribute lag」と「other total contribute lag」がともに有意に支出に影響を与えていること，係数がほぼ同じ値であることが確認された．また，決定係数である Conditional R^2 を確認すると一定の適合率があることが確認された．この結果は，LLM が人間と同じように自身と他者の行動を踏まえ

て支出額を決めていることを意味する.

3.4.支出分布比較

LLM と Human のラウンド 1 およびラウンド 10 の公共財への支出分布を図 3 に示す. ラウンド 1 を確認すると，LLM も Human も 10,15,20 のような切りの良い支出を選択していることが確認された．また LLM も Human も支出 10 が一番多く，分布も同数であった．一方，Human は支出 20 が 17 人に対して，LLM は 1 人と，支出 20 の分布を LLM は十分に再現できていなかった．次にラウンド 10 を確認すると，LLM の支出 20 は増加しているが Human の分布を再現できていないこと, 支出 10 の分布はラウンド 1 と同様に LLM も Human もほぼ同数であることが確認された.

4.考察

4.1. LLM は被験者を代替できるのか？

繰り返し公共財ゲームにおいて LLM は被験者を代替できると考える. 群間比較において支出のベースラインの違いが確認された点や支出分布の支出 20 が再現できていない点という 2 つの課題は存在するが, 支出額決定ロジック比較では人間と同様のロジックで LLM が動作していることが明らかになったため被験者を代替できる可能性があると考える.

4.2. 公共財への支出のベースラインが LLM で再現できなかったのはなぜか？

本研究では LLM に性別や年齢，性格などのペルソナ情報を設定しなかったが，それが影響して公共財への支出のベースラインが下がったと考えられる. ペルソナを LLM に与えた場合の動作は今後研究で確認したいと考えている.

4.3. 支出分布の支出 20 が再現できなかったのはなぜか？

LLM のほうがリスクを考え小刻みに支出を決定しているため，支出 20 のような大胆な行動があまり観測されないのではないかと考える．LLM に「ダイナミックに支出を決定して」のような指示を与えることで，より人間のような動作になる可能性があると考える．この点についても今後の研究で確認したいと考えている.

4.4. LLM で被験者を代替する利点

まず，実験費用面で LLM は圧倒的に費用がかからないことが利点である．今回の LLM での検証にかかった費用は約 720 円であり，Yahoo クラウドソーシングを活用した人間での実験の約 10 分の 1 の費用で実験できた．ラボ実験の場合はさらに費用が掛かることが想定されるため，LLM は費用面で圧倒的に利点がある.

次に，実験準備時間面で LLM は圧倒的に準備の手間がかからないことが利点である. LLM での検証の場合はプログラムを実装するだけでよく，サーバ準備や被験者募集などの手間がかからないため Yahoo クラウドソーシングを活用した人間での実験の約 4 分の 1 の

時間である約 5 時間で準備が完了した．また，LLM の場合は倫理審査が不要という点も準備面では大きな利点と考える．

次に，行動の理由を把握できる点も LLM の利点と考えられる．LLM の場合は都度行動の理由を出力させることができる．また，実験後に行動の理由を出力することも可能である．人間での実験の場合は被験者の手間を考えると都度行動の理由を聞くことは現実的ではない．また，実験中に理由を確認するとそれがバイアスになり以降の行動が変化するリスクも考えられる．そのため人間での実験の場合は細かく理由を確認することは現実的ではないが LLM では細かく理由を確認できる点は大きな利点と考える．

最後に，続きから実験を再開できる点も LLM の利点と考える．LLM の場合，本実験のラウンド 11～20 を後日計測するようなことが可能である．

5.おわりに

本研究では経済実験の被験者を LLM で代替可能なのかを検討するために，繰り返し公共財ゲームを題材に人間と LLM の行動を比較した．結果として，公共財への支出のベースラインや支出 20 の分布を LLM では再現できていないという課題はあるものの，支出決定ロジックが人間と LLM で同様の傾向を示していることより，人間の意思決定を再現できていると考えられる．よって LLM が繰り返し公共財ゲームにおける被験者を代替できる可能性があることが示唆された．今後は本研究で明らかになった課題の解決方法を探求したいと考える．また，繰り返し公共財ゲームの途中で介入があるような実験を LLM で再現できるのかについても検討したいと考える．

引用文献

S. Phelps and Y.I. Russell, 2024. The Machine Psychology of Cooperation: Can GPT models operationalise prompts for altruism, cooperation, competitiveness and selfishness in economic games? arXiv. https://arxiv.org/abs/2305.07970

G. Aher, R.I. Arriaga and A.T. Kalai, 2023. Using large language models to simulate multiple humans and replicate human subject studies. Proceedings of the 40th International Conference on Machine 17, 337-371.

J. Huang, E.J. Li, M.H. Lam, T.Liang, W.Wang, Y.Yuan, W.Jiao, X.Wang, Z.Tu and M.R. Lyu, 2024. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. arXiv. https://arxiv.org/abs/2403.11807

北代絢大，鶴崎祐大，深澤祐援，西野成昭, 2023. 最後通牒ゲームの大規模言語モデルを用いたシミュレーション―経済実験における新手法確立に向けて―. 行動経済学会第 17 回大会.