# 大規模言語言語モデルを用いたエージェントベース進化モデルにおける形質表現の拡張

鈴木麗璽 [1] 浅野誉子 [1] 有田隆也 [1]
[1] 名古屋大学
reiji@nagoya-u.jp asano.takako.u8@s.mail.nagoya-u.ac.jp arita@nagoya-u.jp

## 概要

大規模言語モデル（LLM）の豊かな言語的表現力を活用し，個体間相互作用に関するエージェントベース進化モデルの形質表現を拡張することで，リアルで複雑な集団の進化ダイナミクスを明らかにすることを目的とした 2 つの研究事例を概説する．一つは協力行動に関する多様な性格特性の進化であり，自然言語で記述した高次の心理・認知特性とその行動を LLM を用いて進化モデルに盛り込む試みである．もう一つは，LLM を用いて単語に関連する話し言葉を生成することで，無限に生じる会話トピック選好性の文化進化を表現する試みである．これらを通して，LLM による言語表現に基づく多様で複雑な形質進化から示唆される社会進化ダイナミクスについて論ずる．

## 1 はじめに

ChatGPT 等の大規模言語モデル（以下 LLM）は人と AI との関わり方を急速に改変し知能や意識とは何かという人間の本質にかかわる問題に問いを投げかけている [1]．この潮流において重要となるのは生成モデルに基づく主体間の相互作用 [2] や人が混在する社会の進化ダイナミクスの理解である．

関連する最近の研究として，LLM の認知機能（心の理論 [3]，メタ認知 [4]，ゲーム論的環境における行動と学習 [5, 6]，ビッグファイブ性格特性 [7] など）の理解に関する取り組みが多数行われている．例えば Phelps と Russell は，GPT-3.5 が，社会的ジレンマにおける競争的，利他的，利己的，混合的な動機の自然言語記述を運用する能力を分析した [6]．その結果，LLM は利他主義と利己主義の自然言語記述を解釈でき，ある程度適切に協力行動に反映させることができるが，限界もあることが示された．また，LLM の進化的アルゴリズムへの応用の取り組みも提案されている [8, 9, 10, 11]．例えば，LLM を突然変異や交叉の演算子として利用し，進化計算に創造性とオープンエンド性をもたらす研究がある [10, 11]．Meyerson らは，いくつかのパターンを親として LLM に入力し，関連する新しいパターンを子孫として生成する，Few-shot プロンプトに基づく言語モデル交叉を提案した [10]．バイナリビット列，文章，方程式，テキストから画像へのプロンプト，Python コードの進化に成功している．

一方，社会集団の進化に対するモデルアプローチでは，従来，進化ゲーム論等の数理生物学や，計算社会学・エージェントベースモデル（以下 ABM）等の計算論的手法を用いた抽象的な枠組みに基づいて議論がされ，普遍的な知見がもたらされてきた．しかし，前述のように大規模言語モデルが生み出すような，多様な言語表現に基づく行動規則や，無限に生じる語彙のような自然言語の持つ複雑さをモデルに盛り込むことは容易でなかったといえる．

本研究は，大規模言語モデルの豊かな言語的表現力を活用し，社会における個体間相互作用に関するエージェントベース進化モデルの形質表現を拡張することで，リアルで複雑な集団の進化ダイナミクスを明らかにすることを目的とする．その初期的取り組みとしての 2 つの研究事例について報告する．一つは大規模言語モデルを用いた協力行動に関する多様な性格特性の進化であり，従来のシンプルな数理計算モデルでは表現が容易でなかった高次の心理・認知特性を言語的に記述し，その行動を LLM を用いて生成することで進化モデルに盛り込む試みである．もう一つは，LLM を用いて単語に関連する話し言葉を生成することで，無限に生じる会話トピック選好性の文化進化を表現する試みである．これらを通して，LLM による言語表現に基づく多様で複雑な形質進化から示唆される社会進化ダイナミクスについて論ずる．

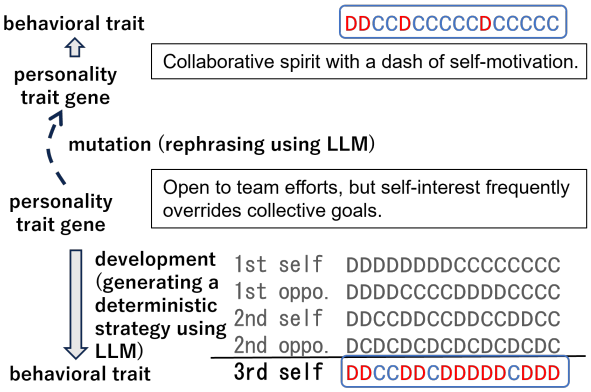



**図 1** LLM を用いた性格特性遺伝子からの行動形質の生成と突然変異操作.

## 2 大規模言語モデルを用いた協力行動に関する性格特性の進化

人間の意図や性格，個性，嗜好のような，行動に複雑に影響し無限の広がりを持つ高次の特徴を直接 LLM を用いて表現した進化モデル構築の試みとして，次のようなゲーム論的状況設定における自然言語で記述された遺伝的性格特性の進化モデルを構築した．詳細は [12] を参照されたい．

$N$ 個体のエージェントの集団を考える．図 1 に示すように，各エージェントは協力的や裏切り的な性格特性に関する十数語から成る英文を性格特性遺伝子として持ち，これに基づいて行動する．その遺伝子から記憶長 4 の決定論的戦略である行動形質を決定するためにチャット型 LLM を用いる．LLM に与えるプロンプトとして，対象個体の性格特性，繰り返し囚人のジレンマゲームの状況設定と利得，対象個体と対戦相手の過去 2 回の行動履歴，および，次のラウンドの行動を答えさせる指示を与える．この方法を用いて可能なすべての履歴（$2^4 = 16$ 個）対する応答を得ることで行動形質を得る．単純化と計算コストの低減のため，上記の行動形質は一意な性格形質遺伝子に対して一度だけ決定，保存され，以後同じ遺伝子が集団に出現する場合には既存の行動形質が使用される．

各組み合わせ $K$ ラウンドの繰り返し対戦からなる総当たり戦における平均利得を適応度とする．対戦時はある確率 $p_n$ でエージェントが意図した行動と反対の行動をとるノイズを導入する．最初のラウンドでは行動はランダムに生成された履歴に基づいて決定される．適応度に比例したルーレット選択に基づく $G$ 世代にわたる進化実験を行う．子個体の生成において，各個体についてある確率 $p_m$ で突然

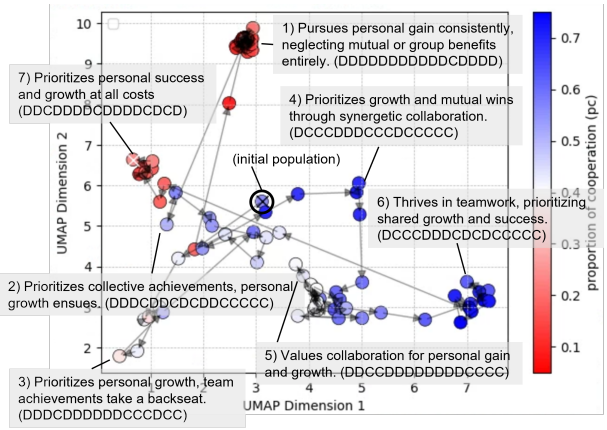



**図 2** 協力・裏切りの入れ替わりが典型的に生じた 1 試行における，10 世代ごとにおける支配的な性格特性遺伝子平均の 2 次元潜在空間上の遷移．色はその世代の協力割合を示す.

変異が起きる．その際，LLM に親の性格特性遺伝子を協力的なもの，もしくは，利己的なものに少し言い換えるように指示し（図 2(右)），その出力文を子孫の遺伝子として採用する．

$N = 30$, $K = 20$, $M$=10, $p_m = 0.05$, $p_n = 0.05$, $G = 1000$，ゲームの利得 $R$=4, $T$=5, $S$=0, $P$=1 を採用した．Huggingface 上で公開されている比較的小規模なチャット型の言語モデル（TheBloke/Llama-2-13b-Chat-GPTQ[1)]）を用いた．

図 2 は，協力的な傾向と裏切り的な傾向とが入れ替わる進化シナリオが明瞭に観察された 1 試行において，2 次元潜在空間における 10 世代ごとの性格特性遺伝子の分布と推移を示したものである．各世代について，SentenceTransformer を用いて全ての性格特性遺伝子をベクトル化し，さらに次元削減アルゴリズム UMAP[13] を用いて 2 次元ベクトルに圧縮した．その後，10 世代ごとの平均ベクトルを 2 次元平面上にプロットした．点の色は対応する世代の協力割合（pc）を示し，いくつかの特徴的な世代における支配的な遺伝子が示されている．

性格特性空間は，集団の性格特性遺伝子と行動形質が協力的なものと利己的なものの間で変動する様子を示している．はじめに，集団は図中左中央から中央上部に向かって利己的な性格特性へと進化した．支配的な性格特性（1:“Pursues personal gain consistently, neglecting mutual or group benefits entirely.（一貫して個人的利益を追求し，相互利益や集団利益をまったく無視する）”）は，この段階ではほぼ

1) https://huggingface.co/TheBloke/Llama-2-13B-chat-GPTQ

完全に裏切り戦略であった．しばらくすると，何度か裏切り集団の侵入を受けつつも徐々に協力的傾向が高まり（2〜5），最終的に最も協力的な集団（6：“Thrives in teamwork, prioritizing shared growth and success.（チームワークに努め，成長と成功を共有することを優先する.）”）に進化し，右下に移動した．しかし，ほぼ完全な裏切り的性格の侵入が，集団を中央左へと導いた．一般的に，集団は裏切り的から協力的なものまで，徐々に性格特性が変化しながら進化した．

全体として，上記のような協力的性格と裏切り的性格の入れ替わりが繰り返し生じる過程が確認された．この挙動は，行動形質をビット列とみなして遺伝子として，各ビットの確率的な反転で表現した対称実験と比べ，より裏切り集団の停滞期が長い傾向があり，裏切り的な特性のほうが多様な表現が生じる傾向が示唆された．また，集団中の性格特性遺伝子中に頻繁に表れた単語のうち，協力行動をよりもたらすものとして “gently, fosters, establishes, harmony”や，より頻繁に相互裏切りや裏切り成功をもたらすものとして，“trampling, trumps, disregard, blatant and skepticism”といった利己的，投機的傾向に関する単語があることなどが分かり，性格特性遺伝子が表す言葉の意味が行動反映されつつ集団の進化が生じていることが示唆された．

総じて，自然言語で記述された性格特性についてLLM を用いて行動傾向を生成し，それに基づいて協力行動を進化させることが可能であることが示された．

## 3　多様な会話トピック選好性に基づく文化進化モデル

人間社会において繰り広げられる雑談や LLM エージェント同士の対話において，トピックに関する遺伝的選好性（個性）が相互作用にもたらす影響と，トピックの伝播に基づく文化進化を理解するため，LLM を発話の生成に用いることで多様な発話表現と膨大なトピックが創発しうるモデルを次のように構築した．同様だがチャット版でない LLM を採用した初期のモデルの詳細については [14] を参照されたい．

集団における社会的な近さを表現した $W \times W$ の 2 次元空間上に $N$ 体の個体が存在し，各個体は会話トピックの選好性を表す単語であり不変の遺伝形質（個性）と他者から受け取る複数の文化形質を持つ（図 3）．各個体は，自身の持つ形質に関連する日本

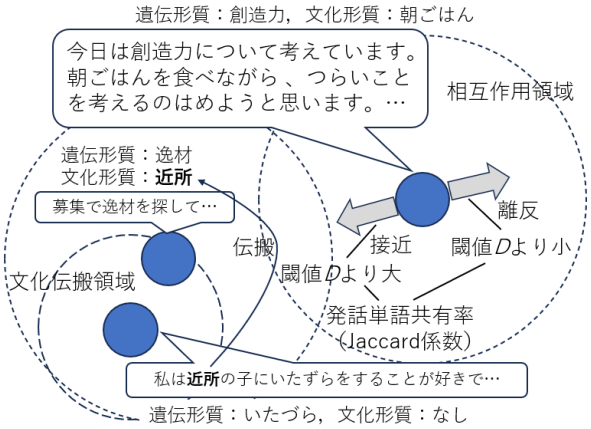



**図 3**　会話文の生成と移動，トピック伝搬．

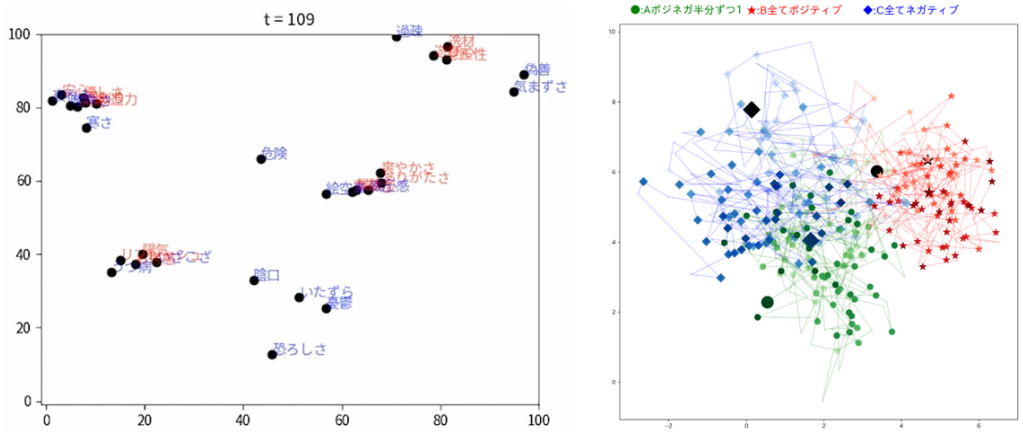
語の発話文を “あなたは、今 (遺伝・文化形質の列挙) という気持ちを持っています。文章内に必ず、(遺伝・文化形質の列挙) という単語を入れた発話文を作ってください。発話文は 200 字以下出力は「承知しました。(改行)」で書き始め、発話文以外のことは出力しないでください。”というプロンプトによって LLM[2] に指示することで生成する．各ステップにおいて確率 $p_r$ で再生成し，発話文を更新する．

各ステップにおいて，各個体は半径 $R_i$ 以内の他個体に対して，確率 $p_t$ で MeCab で単語に分化した発話文について，自身と相手の発話文の単語集合間の Jaccard 係数を測り，閾値 D を超えた（以下の）場合にエージェント間の距離に反比例した引力（斥力）が発生する．すべての他個体からの力の合力方向に一定距離 $V$ だけ進み，他個体が存在しない場合はランダムな方向に進む．これは，現在話している内容が（表面的に）近い他者に興味を持ち，より親密に，積極的に会話に参加するふるまいを表現している．

加えて，ごく近い半径 $R_t(¡R_i)$ 以内に存在する各他個体の発話文に関して，確率 $p_t$ でキーフレーズ抽出処理ライブラリである pke を用いてキーフレーズを抽出し，自身の文化形質に加える．ただし，文化形質を保持する最大数 $L$ を超える場合は最も古いものを破棄する．その上で，現在の遺伝・文化形質を用いて発話文を再生成し更新する．これは，親密な会話によって他者の発話から新たなトピックに興味を持ち，それについて話し出す様子を表現している．

$W$=100，$N$=30，$V$=5，$D$=0.12，$R_i$=20，$R_t$=5，$L$=5 の条件で 200 ステップ実験を行った．各個体の個性

2）https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b-instruct




(a) 遺伝形質（個性）のみ（文化進化なし）の場合の相互作用傾向

(b) 文化形質進化あり，遺伝形質の極性3条件における発話文の平均ベクトルの推移

**図 4**　(a) 遺伝形質（個性）のみ（文化進化なし）の場合の相互作用傾向．(b) 文化形質進化あり，遺伝形質の極性 3 条件における発話文の平均ベクトルの推移．

である遺伝形質の特徴が集団行動に与える影響に注目するため，東北大学 乾・岡崎研究室の日本語評価極性辞書（名詞編）を用いてポジティブ・ネガティブと判定される任意の単語，全員ポジティブ語のみ，ネガティブ語のみ，ポジティブ・ネガティブ半分ずつをそれぞれ各個体の遺伝形質に割り当てるの 3 種の条件を想定した．

予備実験として，単語極性は半分ずつ，文化形質の伝播なし（$p_t$=0.0），再生成なし（$p_r$=0.0）の条件で実験したところ，集団全体として複数の小グループが離散集合する動的な傾向が観察された（図 4）．また，ポジティブ語を持つ個体は話題や話しぶりを共有しやすくグループを形成し維持しがちである一方，ネガティブ語を持つ個体は他人にとって受け入れやすさが異なり，グループから離脱して単独でさまよう傾向があることが分かった．

次に，発話の再生成と文化形質の伝播を有効（$p_r$=0.3，$p_t$=0.4）にして実験したところ，単語極性半分ずつの設定で上記と同様の傾向が確認されたが，全体としてはより動的になり，各個体がより多くの他個体と接して多様な個人間の相互作用が促進された．図 4 は単語極性の 3 種の設定それぞれにおいて実験した際，各ステップにおける各個体の発話文を SentenseTransformer でベクトル化し，UMAP で 2 次元に圧縮したものの集団の平均をプロットしたものである．同図から，ポジティブ語のみ，ネガティブ語のみの集団とは異なる位置にポジティブ・ネガティブ半分ずつの集団が位置しており，極性が異なる個体間の相互作用によって新たな語彙や発話内容が創発しうることが示唆された．

以上のように，各個人が言語的に持つ会話に関する個性が集団全体の相互作用に影響したり，文化の伝播で無数に生じる発話の多様性に影響しうることが示唆された．

## 4　おわりに

大規模言語モデルを活用することで，自然言語を用いたリアルで豊かな形質表現をエージェントベース進化モデルに盛り込む 2 つの試みについて概説した．性格特性の進化モデルでは，協力的裏切り的な性格特性の言語的記述が侵入を繰り返す進化ダイナミクスが観察された．本手法は，問題設定が言語で記述できればあらゆる状況において性格特性から行動を生成可能であり，より多様な環境や複数並列して存在する環境における高次の人間特性の進化を議論できる可能性があると考える．会話トピック選好性に関する文化進化モデルでは，単語で表現され個性として遺伝的に保持するトピックの極性が個体自身や集団のダイナミクスに影響しうることや，近接個体からのトピック伝播に基づく文化進化がより動的な相互作用をもたらしたり，多様な遺伝的トピックの存在が新奇な文化進化をもたらしうることが示唆された．

このような LLM の活用により，従来は現実世界よりも単純であった進化モデルを現実世界と同じくらい複雑にでき，多様な形質の進化シナリオを議論することが可能になるといえる．本論文で概説した試みはこの方向への第一歩となると考えている．

## 謝辞

本研究の一部は，JSPS 課題設定による先導的人文学・社会科学研究推進事業 JPJS00122674991, JSPS 科研費 JP21K12058 の支援を受けた.

## 参考文献

[1] Arend Hintze. ChatGPT Believes It Is Conscious. **arXiv e-prints**, arXiv:2304.12898, March 2023.

[2] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative Agents: Interactive Simulacra of Human Behavior. **arXiv e-prints**, arXiv:2304.03442, April 2023.

[3] Shima Rahimi Moghaddam and Christopher J. Honey. Boosting Theory-of-Mind Performance in Large Language Models via Prompting. **arXiv e-prints**, arXiv:2304.11490, April 2023.

[4] Yuqing Wang and Yun Zhao. Metacognitive Prompting Improves Understanding in Large Language Models. **arXiv e-prints**, arXiv:2308.05342, August 2023.

[5] Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing Repeated Games with Large Language Models. **arXiv e-prints**, arXiv:2305.16867, May 2023.

[6] Steve Phelps and Yvan I. Russell. Investigating Emergent Goal-Like Behaviour in Large Language Models Using Experimental Economics. **arXiv e-prints**, arXiv:2305.07970, May 2023.

[7] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality Traits in Large Language Models. **arXiv e-prints**, arXiv:2307.00184, June 2023.

[8] Benjamín Machín, Sergio Nesmachnow, and Jamal Toutouh. Evolutionary Latent Space Search for Driving Human Portrait Generation. **arXiv e-prints**, arXiv:2204.11887, April 2022.

[9] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-Referential Self-Improvement Via Prompt Evolution. **arXiv e-prints**, arxiv.2309.16797, 2023.

[10] Elliot Meyerson, Mark J. Nelson, Herbie Bradley, Arash Moradi, Amy K. Hoover, and Joel Lehman. Language Model Crossover: Variation through Few-Shot Prompting. **arXiv e-prints**, arXiv:2302.12170, February 2023.

[11] Joel Lehman, Jonathan Gordon, Shawn Jain, Kamal Ndousse, Cathy Yeh, and Kenneth O. Stanley. Evolution through Large Models. **arXiv e-prints**, arXiv:2206.08896, June 2022.

[12] 鈴木麗璽, 有田隆也. 大規模言語モデルを用いた協力行動に関する性格特性の進化モデル. 第 8 回人工生命研究会資料, 2023.

[13] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform Manifold Approximation and Projection. **Journal of Open Source Software**, Vol. 3, No. 29, p. 861, 2018.

[14] 浅野誉子, 鈴木麗璽, 有田隆也. 生成モデルに基づき雑談するエージェントの会話トピック選好性に関する文化進化. 第 37 回人工知能学会全国大会論文集, Vol. JSAI2023, 4H3OS6b04, 2023.