LLMエージェントの集団インタラクションによる性格の分化
Personality Differentiation through Collective Interaction of LLM Agents

高田 亮介 †，升森 敦士 †，池上 高志 †
Ryosuke Takata, Atsushi Masumori, Takashi Ikegami
† 東京大学
University of Tokyo
takata@sacral.c.u-tokyo.ac.jp

概要

本研究では，初期状態やパラメータが同じ複数のLLM エージェントが仮想空間上を移動しながらコミュニケーションするシミュレーション実験を行った．実験の結果，同一のLLM エージェントであるにもかかわらず，エージェントの記憶や周囲へのメッセージ，行動パターン，性格が分化した．このことから，LLM エージェントの個性や性格は，エージェントが空間を移動することで生まれる集団の中でのインタラクションを通して創発し得ることが示唆された．

キーワード：Large Language Model，集団，性格

1. はじめに

GPT-4 (Achiam et al., 2023) などの大規模言語モデル（Large Language Model, LLM）や, ChatGPT などのユーザインタフェースの登場により，人間の生活の一部にLLM が介入するようになった．人間と会話しているかのように自然な会話が実現できるLLM によって，人間の言語的なインタラクションと，そこから創発する個性や社会性などを構成論的に理解することが可能となった．近年では，LLM は人間のユーザとのやり取りだけでなく，ロボットなどの機械における身体性に立脚した運動生成 (Zhang et al., 2023)(Yoshida et al., 2023) や，LLM によって実装されたマルチエージェントシミュレーションによって社会性を創発させる研究 (Park et al., 2023)(Qian et al., 2023) にも使われている．特に Park et al. (2023) の Generative Agents の研究は，25 体のLLM ベースのエージェントを仮想空間に配置しシミュレーションした結果，自律的に各エージェントが日常的な計画を立てたり，自発的にバレンタインパーティを企画するなど，LLM エージェントの集団から社会的な行動が出現した示唆的な研究である．しかし，この Generative Agents をはじめとするこれまでの研究では，LLM にあらかじめ性格などの個性を人間が与えており，個性自体がどのように創発するかは明らかになっていなかった．

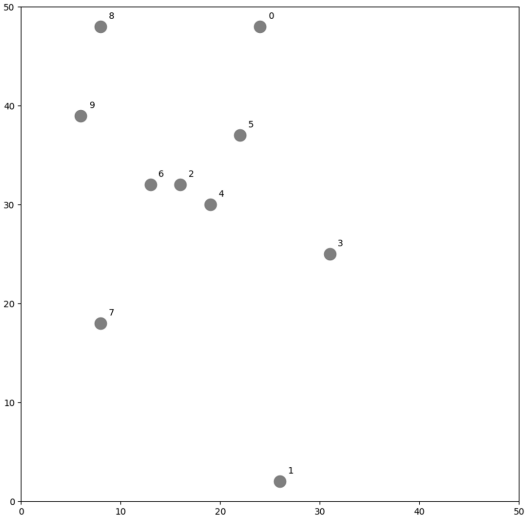

図 1　シミュレーション空間．50x50 の 2 次元グリッド空間に 10 体の LLM エージェントがいる．エージェントの初期配置はランダムに決定する．空間は周期的境界条件を持っている．

本研究では, 個性を持たないホモジーニアスなLLM エージェントの集団において，周囲のLLM エージェントとのインタラクションを通して個性や性格を分化させ得るか検証することを目的とする．同じLLM のパラメータを持ったホモジーニアスなエージェント集団を作成し，2 次元仮想空間上を移動させるシミュレーションを通して，各エージェントの行動特性や性格がどのように創発するかを分析した．

2. LLM エージェントとシミュレーション空間

本シミュレーションでは，50x50 の 2 次元グリッド空間内に同じパラメータを持つ LLM ベースの 10 体のエージェントを配置し（図 1），100 ステップのシミュレーションを行った．環境にはエージェント以外に何も配置せず，エージェントに対しては特にタスク

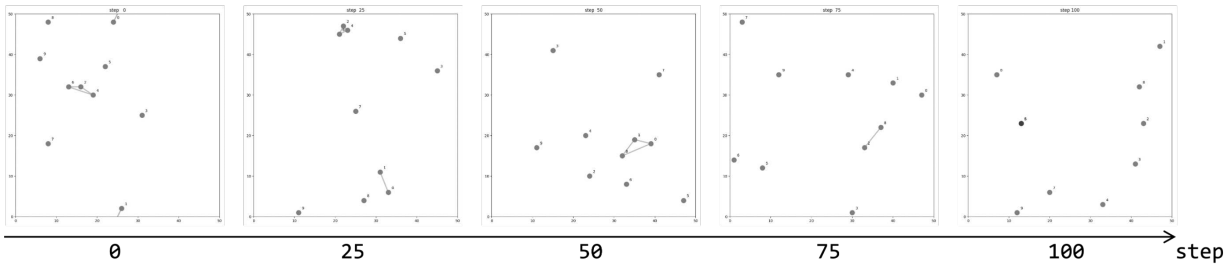

図 2　100 ステップのシミュレーションの結果，LLM エージェントの移動による位置変化．線で繋がっているエージェントはメッセージ交換の範囲内にいることを表す．メッセージ交換するエージェントのクラスタは動的に変化していることがわかる．

表 1　LLM のパラメータ一覧

パラメータ	値
モデル	Llama2-7b-chat-hf
温度	0.7
最大トークン数	256
top-p	0.95
top-k	40

を与えずそれぞれ自由に会話しながら動き回るように指示した．エージェントは各ステップで，(1) 周囲のエージェントへのメッセージ送信，(2) 自身の最近の活動を要約した状況的記憶の保存，(3) 次の行動（どこへ移動するか）の選択，という 3 つのアクションを行う．この 3 つのアクションを生成するために，LLM を使用した．

LLM エージェントは，周囲の距離 5 以内の LLM エージェントとメッセージを交換する．このとき当該範囲内に複数のエージェントが存在した場合は，その全てのエージェントとメッセージを交換する．また，エージェントの記憶は, LLM によって生成される自然言語の出力をそのまま保持するように設計した．エージェントの行動は，グリッド空間における上下左右の移動または静止のいずれかを LLM に選択させる形式で生成させた．ここで，LLM のプロンプトは，LLM エージェントの現状態（エージェント名，その時点での座標，記憶），周囲から受け取ったメッセージ，上述の (1) から (3) に対応する指示文によって構成した．

今回，LLM には Meta 社がリリースした Llama 2(Touvron et al., 2023) を使用した．Llama 2 は 2 兆個のトークンと 4096 のコンテキスト長で事前学習されている．ここでは，70 億パラメータのモデルを用いており，LLM のパラメータは表 1 の通りとした．

3. シミュレーション結果と分析

3.1 行動の変容

シミュレーションの結果, LLM エージェントは移動する行動を多く生成し，移動した先にいる LLM エージェントとメッセージを交換した．この LLM エージェントの移動に伴って，メッセージを交換するエージェント数やメッセージ内容は動的に変化した（図 2）．

生成された行動に注目すると，同一の LLM を用いているにもかかわらず，エージェントの “静止” などの行動の生成頻度に違いが見られ，それによってエージェントごとの移動パターンに差異が生じていた．さらに，この差異は他のエージェントとメッセージを交換した経験の有無に影響されていることがわかった．このことから，エージェントが周囲のエージェントとの相互作用によって自らの行動を変容させたことが示唆された．

3.2 記憶とメッセージの生成

LLM によって自律的に生成させたエージェントの記憶の全体的な変化を分析した．自然言語で記述された記憶とメッセージをそれぞれベクトル化し，その多次元ベクトルに対して 2 次元への埋め込み表現を行う UMAP(McInnes et al., 2018) を用いた．これにより，言語的に近い意味を持つ文章は近い位置にプロットされるため，シミュレーションを通して出現したトピックの全体像を捉えることができる．

分析の結果，LLM エージェントの保持する記憶はエージェントごとに全く異なっていたのに対し，LLM エージェントの発したメッセージは複数エージェントの間で近い内容に変化していた（図 3）．LLM エージェントの保持する記憶の情報流は内部状態として各エージェントの内部で閉じているのに対し，LLM エージェントの外部と送受信されるメッセージの情報流は開かれている．この情報流の性質によって，メッセージの UMAP はエージェント集団ごとに多様化し，

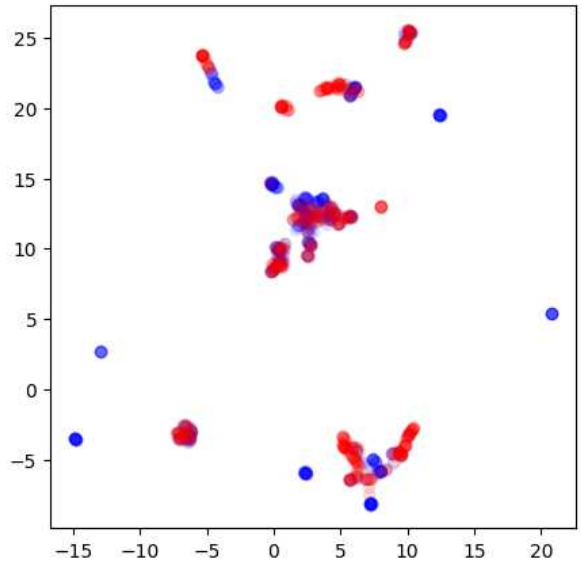

図 3　メッセージの UMAP．全てのエージェントが生成したメッセージをベクトルに変換し，その相対的な距離を保ったまま 2 次元に埋め込み圧縮した．青いプロットは前半ステップ，赤いプロットは後半ステップ．後半ステップでは前半ステップに比べてプロットのばらつきが少なくなっていることがわかる．

かつ集団内部では一様化したと考えられる．すなわち，開かれた情報源としてのメッセージは，エージェントが集団を形成したときに自己組織化しやすいのに対し，閉じた情報源としての記憶は自己組織化しにくいことが示唆された．

3.3　性格の分化

人間に対して特定の質問を行い性格を評価する方法と同様に，LLM に対しても特定の質問によって性格を評価する方法が有効である可能性が示唆されている (Jiang et al., 2024)．本研究では，MBTI 性格テストを用いた性格評価分析 (Boyle, 1995) により，初期状態ではほとんど同一の性格であったエージェントが，集団での相互作用を通じて異なる性格タイプに分化したことが認められた（表 2）．シミュレーション開始時（0 ステップ時点）とシミュレーション終了時（100 ステップ時点）のそれぞれの記憶を持つ 10 体の LLM エージェントに対して，Pan and Zeng (2023) で提案されている質問項目を用いて MBTI 性格評価を行った．0 ステップ時点では 10 体の LLM エージェントの差異は名前と初期位置だけであるためほとんど同じ性格と評価されたが，100 ステップ時点では 5 つの異なる性格評価に分かれた．LLM はモデルごとに性格が異なることが分かっており，本研究で初期状態の性格として評価された INFJ は，今回使用した Llama 2 におけるデフォルトの性格タイプとして知られている (Pan and Zeng, 2023)．一方で最終的に分化した性格は全て初期状態とは異なっており，大別して，リーダー的役割とフォロワー的役割に分類された．このように，同一の LLM エージェントにおいても，集団でのインタラクションを通して自律的に異なる性格を獲得したことは，性格などの個性は集団から創発し得ることを示唆している．

表 2　10 体の LLM エージェントにおける初期状態（0 ステップ時点）と最終状態（100 ステップ時点）の MBTI 性格テスト結果

Agent	MBTI Type	
	step 0	step 100
agent0	INFJ	ESFJ
agent1	INFJ	ISTJ
agent2	INFJ	ISTJ
agent3	INFJ	ENTJ
agent4	INFJ	ISTJ
agent5	INFJ	ISTJ
agent6	INFJ	ESTJ
agent7	INFJ	ENTJ
agent8	INFJ	ENTJ
agent9	INTJ	ISFJ

4.　まとめ

本研究では，仮想空間上を移動する LLM エージェントの集団シミュレーションを通して，集団インタラクションから創発する個性について議論した．分析結果から，各エージェントがメッセージを生成する際に，周囲のエージェントとの相互作用によって集団ごとに近い内容のメッセージを生成し自己組織化することが示唆された．また，全ての LLM エージェントは同一パラメータを持ったホモジーニアスな個体集団であるにもかかわらず，行動や性格が分化した．以上のことから，性格などの個性は集団インタラクションを通して創発することが示唆された．

今後の研究では，エージェント数やステップ数を増やし，より大規模なシミュレーションを行うことで，集団サイズと個性の関係について検討していく．また，空間内に壁などの障害物を配置することで，空間性と個性についての議論を行うことも可能となる．これらのシミュレーションに対して LLM エージェントの行動を意味論的に分析することで，集団を作る他者や環境とのインタラクションや，その中で個性や社会性を創発させる機序の解明に寄与し得る．

文　献

Achiam, J. et al. (2023). GPT-4 Technical Report. arXiv:2303.08774.

Zhang, Y. et al. (2023). MotionGPT: Finetuned LLMs are General-Purpose Motion Generators. arXiv:2306.10900.

Yoshida, T. et al. (2023). From Text to Motion: Grounding GPT-4 in a Humanoid Robot “Alter3”. arXiv:2312.06571.

Park, J. S. et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior, *In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22.

Qian, C. et al. (2023). Communicative Agents for Software Development. arXiv:2307.07924.

Touvron, H. et al. (2023). LLaMA 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

McInnes, L. et al. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426.

Jiang, G. et al. (2024). Evaluating and Inducing Personality in Pre-trained Language Models. Advances in Neural Information Processing Systems, 36.

Boyle (1995). Myers-Briggs Type Indicator (MBTI): Some Psychometric Limitations. Australian Psychologist, 30(1), 71-74.

Pan, K. and Zeng, Y. (2023). Do LLMs Possess a Personality? Making the MBTI Test an Amazing Evaluation for Large Language Models. arXiv:2307.16180.