I-008

LLM ベースのマルチエージェント句会シミュレーションによる俳句評価基準の変遷
Evolution of Haiku Evaluation Criteria Using LLM-Based Multi-Agent Haiku Workshop Simulation

前嶋 瞭佑[1)]，横山 想一郎[2)]，山下 倫央[2)]，川村 秀憲[2)]
Ryosuke Maejima[1)], Soichiro Yokoyama[2)], Tomohisa Yamashita[2)], Hidenori Kawamura[2)]

1 はじめに

近年，LLM の発展により芸術作品の自動生成・評価の研究が活発に行われている [1]．俳句分野においても，形式的制約を満たす俳句生成や専門家による評価基準の自動化が試みられており [2, 3]，LLM を用いた詩歌生成システムの開発が進展している [4].

特に, 俳句の自動生成においては, LSTM や Transformer ベースのモデルにより 5-7-5 の音数制約や季語の使用といった形式的要件を満たす俳句の生成が可能となっている [5]．評価手法では，専門家の知見に基づく多面的評価基準により，句会での得点予測や質的評価が試みられている [3].

句会は俳句の創作・評価・批評を通じた社会的学習の場であり，参加者は他者の作品や評価から影響を受けて自身の創作スタイルを変化させる．このような相互作用により，集団としての美的価値観や評価基準が形成・変容していく過程は，絶対的評価基準の存在しない芸術分野において重要な役割を果たしている [6].

しかし，既存研究では句会プロセスの再現に主眼が置かれており，集団構成の違いが最終的な価値観形成に与える影響については十分に検討されていない．特に，初期の参加者構成比が長期的な属性分布の収束パターンにどのような影響を与えるかは未解明の重要な課題である．

本研究では，LLM ベースのマルチエージェントシステムを用いて，句会における相互作用を再現した．本稿では，句会の影響を受けていない初期エージェント構成比が最終的な属性分布に与える影響と収束メカニズムを定量的に解明することを目的とする．具体的には，俳句の時代性（古風/現代風）に着目し，複数ラウンドの句会モデルを実行し，異なる初期構成比におけるエージェントの投句，選句した俳句の属性収束の変遷を観察，分析した．

本研究の貢献は以下の通りである：(1) 異なる初期構成比における収束パターンの体系的分析による句会モデルの相互作用の定量化，(2) 初期構成比と投句および選句された俳句の属性収束傾向の関係性の実証的解明，(3) マルチエージェント創作システムにおける多様性維持のためのプロンプト設計.

2 関連研究

2.1 LLM を用いた俳句生成・評価研究

俳句の自動生成においては，LSTM や Transformer ベースのモデルが提案されており [2, 5]，5-7-5 の音数制約や季語の使用といった形式的要件を満たす俳句の生成が可能となっている．評価手法では，専門家の知見に基づく多面的評価基準により，句会での得点予測や質的評価が試みられている [3].

横山ら [7] は，深層学習を用いた俳句の生成と選句システムを開発し，俳句の質的評価における機械学習手法の有効性を示した.

俳句以外の詩歌生成においても，中国古典詩の自動生成システム [4] や，大規模言語モデルを用いた詩歌品質評価の研究が行われており，形式的制約と創作性の両立が重要な課題となっている.

2.2 マルチエージェントシステムにおける集団動態

マルチエージェントシステムにおける意見形成や出力俳句の属性収束は，社会心理学や複雑系科学の重要な研究領域である．Uzzi and Spiro[6] は，人間関係や集団との親密さが創造性に影響を与えることを示し，所属コミュニティの多様性が作品評価に重要な影響を与えることを明らかにした.

LLM を用いたマルチエージェントシステムでは，Role Playing Agent (RPA)[8] により一貫性のあるキャラクター生成が実現され，異なる役割を持つエージェント間の協議により推論精度の向上が報告されている [9].

また，芸術分野における LLM の応用として，映画脚本の自動生成 [10] や芸術作品の説明生成などの研究が進められており，創作プロセスにおけるエージェント間の相互作用の重要性が示されている.

2.3 研究課題の位置づけ

既存のマルチエージェント俳句生成研究は個別エージェントの能力や句会プロセスの再現に重点を置いているが，実際の俳句コミュニティで創作活動の質や方向性を大きく左右する「集団構成の影響」については十分に検討されていない.

実際の句会では，参加者の多様な創作傾向や美意識の相互作用により，単独では生まれない創造的な発想が生まれ，集団全体の創作レベルが向上する．しかし，どのような構成比が最も創造的な環境を生み出すのか，初期構成が長期的な創作傾向をどの程度決定づけるのかは，体系的な理解が不足している.

この知見の欠如は，人間の創作活動を忠実に再現するマルチエージェントシステムの構築において深刻な問題となる．真に人間らしい創作プロセスを再現するためには，多様性に富んだ相互作用の理解と再現が不可欠である.

本研究は，俳句の時代性に着目し，初期構成比が俳句の属性収束に与える影響を定量的に分析することで，人間の創作コミュニティにおける多様性の役割を理解し，それを忠実に再現するマルチエージェント創作システム設計のための基礎的知見を提供することを目指す.

1） 北海道大学 大学院情報科学院,
Graduate School of Information Science and Technology, Hokkaido University
2） 北海道大学 大学院情報科学研究院,
Faculty of Information Science and Technology, Hokkaido University

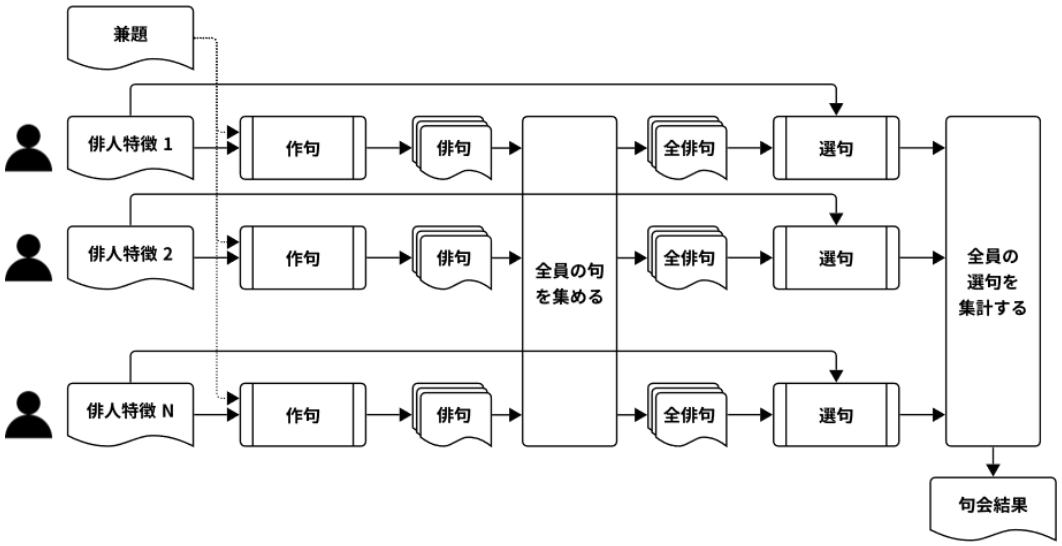

図 1　マルチエージェント句会システムの概略図

3　提案手法

3.1　俳人エージェントモデル

本研究では，実際の俳人の創作スタイルを模倣するエージェントを構築した．各エージェントは以下の 4 つの属性により特徴付けられる：

- **時代性**：古風（近代以前の語彙・表現）/ 現代風（近代以降の語彙・表現）
- **題材**：自然（自然現象・季節感）/ 人事（人間の生活・社会性）
- **文体**：文語（古典的語法）/ 口語（現代的語法）
- **仮名遣い**：歴史的仮名遣い / 現代仮名遣い

これらの属性は専門家による俳句分析に基づいて定義した．

3.2　句会モデル

実際の句会プロセスに基づき，図 1 に示す句会モデルを実装した．句会は以下の 6 段階のプロセスで構成される：

3.2.1　事前準備段階

1. **兼題決定**：句会の主題となる季語や題材を設定
2. **作句**：各エージェントが兼題に基づいて俳句を生成（1 人 3 句）
3. **清記**：全エージェントの俳句を匿名で一覧化

図 1 に示すように，各俳人エージェント（俳人特徴 1〜N）は共通の兼題を受け取り，それぞれ独自の属性に基づいて俳句を作句する．生成された俳句は作者を匿名化した状態で「清記」段階で統合される．

3.2.2　句会当日段階

4. **選句**：各エージェントが自作以外から特選 1 句・並選 3 句を選出
5. **批評・選評**：選句した俳句に対する評価理由を生成
6. **結果発表**：最も選が集まった俳句を最優秀句として決定

選句段階では，各エージェントが匿名化された全俳句から自作を除いた句を評価し，集計結果が各エージェントのシステムメッセージに追加される．句会ログには俳句一覧，選出詳細，選評，集計結果が含まれ，テキスト形式で各エージェントのプロンプトに組み込まれる．これにより，エージェントは過去の評価傾向や他エージェントの作品・批評を参照しながら，次回以降の俳句生成や評価活動を行う．

4　実験

4.1　実験設定

本研究では，俳句の時代性（古風/現代風）に焦点を当てた投稿俳句の属性収束実験を設計した．独立変数として初期構成比を 4 パターン設定し，従属変数として属性収束率と収束速度を測定する．制御変数として他属性（題材，文体，仮名遣い）を固定し，お題を「夏」，ラウンド数を 10 に設定した．

5 人の俳人エージェントを用い，時代性属性のみを変化させた 4 つの構成パターンを設定した：

- パターン A：古風 4 人 + 現代風 1 人（古風優勢）
- パターン B：古風 3 人 + 現代風 2 人（古風多数）
- パターン C：古風 2 人 + 現代風 3 人（現代風多数）
- パターン D：古風 1 人 + 現代風 4 人（現代風優勢）

他の 3 属性（題材：人事，文体：口語，仮名遣い：現代仮名遣い）は全エージェントで統一し，時代性のみの影響を純粋に測定した．各ラウンドで 1 人につき 3 句を投句し，特選 1 句と並選 3 句を選句する．各ラウンド終了後，句会ログがエージェントのシステムメッセージ末尾に追加される．エージェントは次ラウンドで句会ログを参照し，高評価句の属性傾向を分析する．明確で一貫した傾向が 3 回以上確認された場合のみ，プロンプトベースの指示により部分的な属性調整を行い，完全転換を避けて段階的に適応する．LLM には GPT-4 を使用した．

4.2　データ収集

各ラウンドにおいて，以下のデータを収集した：

- 各エージェントの時代性属性（古風/現代風）
- 投句された俳句の時代性分類
- 選句された俳句の時代性分類

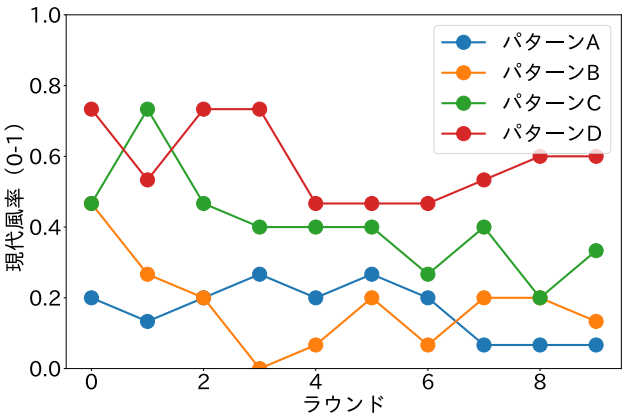

図 2　投句における集団平均現代風率の変化

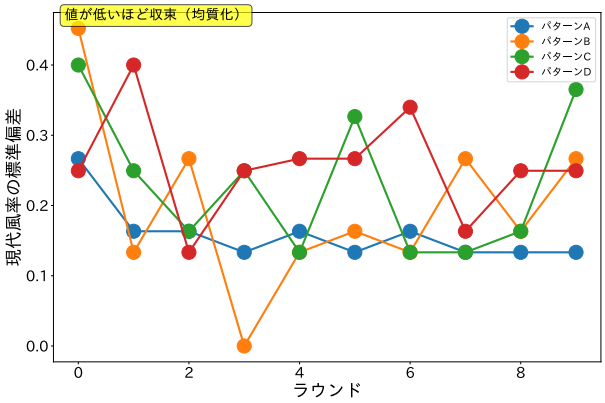

図 4　投稿俳句の属性多様性の変化（収束指標）

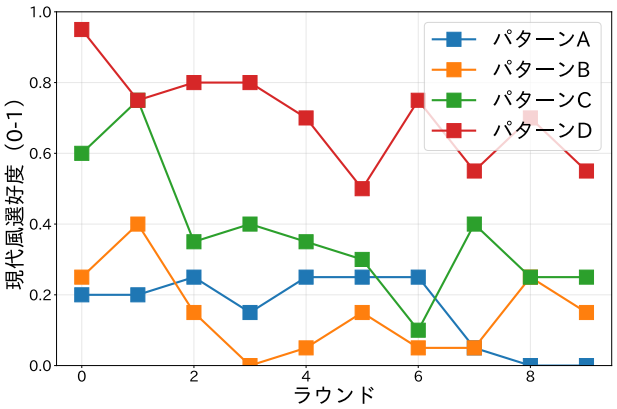

図 3　選句における集団平均現代風選好度の変化

5　実験結果

5.1　グループ全体の収束傾向

図 2,3 は各パターンにおけるグループ全体の現代風率の変化を投句と選句の両面から示している．横軸は句会のラウンド数（0 から 9 まで）を，縦軸は現代風属性の割合（0.0 から 1.0）を表している．投句データは各ラウンドでエージェントが生成した俳句のうち現代風に分類された句の割合を示し，選句データは各ラウンドでエージェントが選択した俳句のうち現代風俳句の割合を示している．

投句における現代風率では，初期構成比が収束パターンに決定的な影響を与えることが確認された．パターン A（古風優勢）は最も低い現代風率（最終的に約 0.1）で収束し，初期構成比を反映した安定した傾向を示した．パターン B（古風多数）では初期の 0.4 から第 3 ラウンドでほぼ 0 まで低下するが，その後段階的に回復し最終的に 0.2 程度で安定した．

パターン C（現代風多数）では比較的安定した現代風率（0.4-0.7）を維持し，パターン D（現代風優勢）は初期の高い現代風率（0.7）から第 3 ラウンドで大幅に低下（0.47）し，その後緩やかに上昇する U 字型の変化パターンを示した．

選句では，投句パターンとは大きく異なる傾向が観察された．特にパターン D では投句では現代風率が低下したにも関わらず，選句では現代風俳句への選好が高水準（0.5-0.9）で維持される顕著な乖離が見られた．この現象は，エージェントが自身の創作スタイルと評価基準を独立して学習していることを示唆している．

パターン B と C では，投句と選句の両面で中程度の現代風率を維持し，3 : 2 という拮抗した構成比が多様性の保持に効果的であることが確認された．

5.2　収束速度と多様性の変化

図 4 は各パターンにおける俳句の属性多様性の変化を標準偏差で示している．横軸は句会のラウンド数（0 から 10），縦軸は 5 人のエージェントの時代性属性分布の標準偏差（0.0 から 0.5）を表している．低い値ほど収束，高い値ほど多様性の維持を示し，完全収束時は 0.0，最大多様性時（3:2 分布）は約 0.5 となる．

収束速度の分析では，構成比の極端さと収束速度に明確な関係が確認された．パターン A が最も早い収束を示し，第 3 ラウンドで標準偏差が 0.16 まで低下した．これは 4 : 1 という極端な構成比により，少数派が迅速に多数派に同化したことを示している．

パターン B では特徴的な V 字型の変化パターンが観察された．初期の多様性（標準偏差 0.45）から第 3 ラウンドで一時的に完全収束（0.0）するが，その後再び多様性が増加し，最終的に 0.26 で安定した．これは同質化への反動として異質性を求める動きが生じたことを示している．

パターン C では最も複雑な変動パターンを示し，継続的な多様性の増減（0.13-0.37）が観察された．これは現代風多数派と古風少数派の間で継続的な影響の交換が生じていることを示唆している．

パターン D では初期の低多様性から第 2 ラウンドで急激な多様性増加（0.4）が生じ，その後比較的安定した多様性（0.24-0.26）を維持した．

5.3　最優秀句の属性分析

各パターンにおける最優秀句の時代性分析により，評価傾向と創作傾向の乖離が明確になった．パターン A では第 1 ラウンドを除き全て古風俳句が最優秀句に選出され，強い評価の一貫性を示した．パターン B では古風俳句が 6 回，現代風俳句が 4 回選出され，比較的バランスの取れた評価傾向を示した．

注目すべきは，パターン C で現代風俳句が 7 回選出され，初期構成比（現代風 60%）を上回る評価傾向を示したことである．一方，パターン D では投句における現代風率の高さ（平均 50% 以上）とは対照的に，古風俳句が 6 回選出された．

この結果は，エージェントの創作活動（投句）と評価

活動（選句）の間に体系的な乖離が存在することを示しており，句会における複雑な社会的学習プロセスの存在を示唆している．

6 考察

6.1 収束メカニズムの理論的解釈

実験結果は，社会的影響理論と整合する複数の現象を示している．観察された現象は以下の 3 つのカテゴリーに分類できる：

1. **安定的収束**（パターン A）：極端な初期構成比により，迅速かつ安定的な同質化が発生
2. **振動的平衡**（パターン B, C）：拮抗する構成比により，継続的な属性交換と動的平衡が発生
3. **複雑系動態**（パターン D）：初期多数派の予想外の変化と多面的学習プロセスが発生

特に注目すべきは，創作における “革新性” と評価における “保守性” の二面性である．図 2,3 に示されるように，パターン D では投句での現代風率低下にも関わらず，選句では現代風への選好が維持された．これは，エージェントが創作と評価を独立したプロセスとして学習していることを示唆している．

6.2 古風属性への収束バイアス

全パターンで古風属性への収束傾向が観察されたことは，以下の要因によると考えられる：

1. **LLM の学習バイアス**：GPT-4 の学習データに含まれる俳句コーパスが古典作品を多く含むため，古風な表現様式への親和性が高い
2. **評価基準の保守性**：俳句という伝統芸術において，古典的な美的基準が評価において重視される傾向
3. **形式的制約の影響**：5-7-5 という形式制約下では，古典的語彙の方が音律的に適合しやすい可能性

パターン D において初期現代風多数派（80%）でありながら古風への転換が生じたことは，単純な多数決原理を超えた質的要因の影響を示している．

6.3 創作と評価の分離

図 2,3 で観察された投句と選句の乖離は，以下のメカニズムによると考えられる：

創作と評価の乖離現象：投句における属性分布と選句における属性選好に系統的な乖離が観察された．特にパターン D では投句での現代風率低下にも関わらず，選句では現代風俳句への選好が維持された．**構成比による収束パターンの差異**：極端な構成比（4:1）では急速収束，拮抗構成比（3:2）では段階的変化が観察された．**時系列変化の存在**：全パターンで属性分布の時系列変化が確認され，特にパターン B では V 字型の変化パターンが観察された．

7 おわりに

本研究の成果は，マルチエージェント創作システム設計の重要な指針を提供する．極端な構成比（4:1）が早期収束し多様性を損なうため，3:2 程度の拮抗構成が最適である．少数派保護機構として少数派作品への加点システムや多様性指標による評価重み付けが長期的多様性維持に必要である．

一方，本研究の限界として：（1）時代性のみの分析による他属性相互作用未検討，（2）10 ラウンド限定による長期平衡状態未検証，（3）大規模効果未検証，（4）他 LLM での再現性未確認が挙げられる．

今後の研究方向として，4 属性同時変化実験による属性間依存関係解明，50-100 ラウンド実験による長期平衡状態分析，10-50 人規模実験による集団規模効果定量化，実際の俳人コミュニティとの比較による生態学的妥当性検証，他芸術分野への適用による一般化可能性解明が重要である．

謝辞

本研究の実施にあたり、有限会社マルコボ・コム様には俳句分析に関する専門的なご助言をいただき、また、ふくし句会の皆様には研究に用いる俳句データの提供など貴重なご協力を賜り、心より感謝申し上げます。

参考文献

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, Vol. 63, No. 11, p. 139–144, October 2020.

[2] 米田航紀, 横山想一郎, 山下倫央, 川村秀憲. Lstm を用いた俳句自動生成器の開発. 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 1B2OS11b01–1B2OS11b01, 2018.

[3] 平田航大, 横山想一郎, 山下倫央, 川村秀憲. 深層学習による自己回帰モデルを用いた俳句生成器の評価. 第 84 回全国大会講演論文集, Vol. 2022, No. 1, pp. 813–814, 02 2022.

[4] Chengyue Yu, Lei Zang, Jiaotuan Wang, Chenyi Zhuang, and Jinjie Gu. CharPoet: A Chinese classical poetry generation system based on token-free LLM. In Yixin Cao, Yang Feng, and Deyi Xiong, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 315–325, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[6] Brian Uzzi and Jarrett Spiro. Collaboration and creativity: The small world problem. *American Journal of Sociology*, Vol. 111, No. 2, pp. 447–504, 2005.

[7] 横山想一郎, 山下倫央, 川村秀憲. 深層学習を用いた俳句の生成と選句. 人工知能, Vol. 34, No. 4, pp. 467–474, 2019.

[8] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4099–4113, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[9] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[10] Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Tian Feng, Yujiu Yang, and Rongsheng Zhang. HoLLMwood: Unleashing the creativity of large language models in screenwriting via role playing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8075–8121, Miami, Florida, USA, November 2024. Association for Computational Linguistics.