Minecraft における大規模言語モデルエージェントの共同作業の実現可能性の調査

廣田 悠希　東中 竜一郎
名古屋大学大学院 情報学研究科
hirota.yuki.a6@s.mail.nagoya-u.ac.jp
higashinaka@i.nagoya-u.ac.jp

概要

近年，Minecraft を用いた共同作業の研究が進んでいるが，エージェントが単独で建築を行う設定や所定の問題解決が主であり，対話しながらの共同建築は十分に検討されていない．また，先行研究である Action-Utterance Model は学習データの不足による限界があった．本研究では，Action-Utterance Model と Creative Agents の枠組みを統合し，高性能な大規模言語モデルによるエージェントを構築する．このエージェントに共同建築タスクを実行させ，生成された庭と対話を評価した結果，一定品質の庭を作成でき，意見の主張も積極的に行える一方で，合意形成に向けた応答が十分でなく，人間と同等の共同作業に到達するには課題が残ることを確認した．

図 1　エージェントが協力して庭を作成する様子．エージェントは Gemini 2.0 Flash に基づいている．エージェントの構築方法については 4.1 節を参照．

1　はじめに

近年，GPT-4 [1] など大規模言語モデル（LLM）の発展により，対話エージェントの性能は大きく向上している [2, 3]．これに伴い，共同作業タスクにおいて人間と協力する対話エージェントの研究も活発化している [4, 5, 6]．

身体性を持つ対話エージェントの共同タスク研究は，費用面の制約から仮想環境で進められることが多く，Minecraft は主要な研究基盤として利用されている [7, 8, 9]．しかし既存研究の多くは，エージェントが単独で建築する設定 [10] や特定問題を解決する設定 [11, 12] が中心で，対話を通じた創造的な共同建築の研究は限定的である．この課題に対し，市川ら [13] は対話を通じて庭を設計する共同建築タスクを提案し，人間データで小規模 LLM をファインチューニングした Action-Utterance Model [14] を提案した．Action-Utterance Model は，各ターンで行動種別を選択し，対応する発話またはブロック操作を生成する逐次的な行動選択フローを採用する．一方で，このモデルは学習データ不足やモデル化の不十分さから性能には限界があった．

一方，テキスト指示に基づく建築タスクでは LLM が顕著な進展を示しており，適切なエージェント設計により創造的な共同作業も実現できる可能性がある．そこで本研究では，最先端の LLM を用いて共同建築エージェントを構築し，エージェント同士が協力して庭を作成できるかを検証する．具体的には，単一指示から高品質な構造物を構築する Creative Agents [10] に対して Action-Utterance Model の逐次的な行動選択フローを導入する．図 1 は，本研究で構築したエージェントが協力して庭を作成する様子を示す．

本研究では，共同建築タスクをエージェント同士で実行し，生成された庭と対話を人手評価する．さらに，同一タスクを実行した人間同士の結果と比較し，改善に向けた知見を得る．

2 関連研究

本研究に関連する分野として，Minecraft で行動するエージェント研究と，LLM を用いた共同タスクエージェント研究が挙げられる．

2.1 Minecraft で行動するエージェント

Minecraft はサンドボックス型の 3D 環境であり，プレイヤーやエージェントが環境と自由に相互作用できることから，研究基盤として広く利用されている [15, 16]．特に建築に焦点を当て，Zhang ら [10] は Creative Agents を提案した．Creative Agents は，テキスト指示に基づいて構造物を建築するエージェントであり，単一指示から構造物全体を構築できる．一方，単一の指示に基づく建築を主に想定しているため，対話を通じたインタラクションや，段階的な建築はできない．

2.2 LLM を用いた共同作業エージェント

Chiu ら [17] は所定のドットを協力して特定する OneCommon タスク [18] を対象に，GPT-4 を用いたエージェントが高い性能を示すことを報告した．

市川ら [13] は，Minecraft において協力して庭を設計・建築する共同建築タスクを提案した．また，小規模な LLM[1)]をファインチューニングすることで，このタスクを実行する Action-Utterance Model [14] を提案した．このモデルは，各ターンで行動種別を選択し，対応する出力を生成する行動選択フローを採用することで，ベースラインより適切な行動・発話を生成することが可能であるが，学習データの不足による性能の限界があった．

3 アプローチ

本研究では，共同建築タスクを実行可能なエージェントを構築するため，Action-Utterance Model と Creative Agents の枠組みを統合する．

Action-Utterance Model は，対話履歴とワールド状態を入力として，次の行動種別を選択し，選択に対応する出力を生成することで，発話とブロック操作を逐次的に扱う．一方，Creative Agents は最先端 LLM により，単一のテキスト指示から高品質な構造物を構築する非対話型のエージェントであり，ブロックの種類や配置の決定は高精度であるが，対話を通じた段階的な共同建築は想定されていない．

本研究では，Creative Agents にプロンプト拡張を通して Action-Utterance Model の行動選択フローを組み込むことで，両者を統合する．具体的には，まず，Action-Utterance Model の行動選択フローである，各ターンで次の行動種別（CHAT, BLOCK, SKIP, FINISH）を選択し，選択に応じて発話生成とブロック操作を切り替えるフローを導入する．また，従来の Creative Agents が用いる単一の指示を，対話履歴とワールド状態を含む文脈入力に置き換える．さらに，ブロック操作に加えて，与えられた文脈に適した発話も生成できるようプロンプトを拡張する．

以上により，Creative Agents の高精度な生成能力と Action-Utterance Model の逐次的な意思決定を統合し，対話を通じて創造的に共同建築できるエージェントを構築する．

4 実験

本節では，3 節の手法で構築したエージェントの性能を評価する実験について述べる．

4.1 実験手順

本研究では，2 体のエージェントで共同建築タスクを実施した．作成された庭と対話内容について，統計量に基づく客観評価とクラウドソーシングによる人手評価で評価した．

エージェントに使用する最先端の LLM として，GPT-4o，Gemini 2.0 Flash，DeepSeek-V3，Llama 3.3 70B Instruct，Qwen3 32B の 5 種類の LLM を用いた．対話はすべて日本語で行い，各エージェントに日本語で発話するよう指示した．

各 LLM に共同建築タスクを 20 回ずつ実行させ，計 100 件の庭と対話を生成・評価した．さらに上限として，人間同士が同タスクを実行して作成した庭と対話も評価に含めた．具体的には，共同建築タスクコーパス [13] から人間の対話をランダムに 20 件抽出した．

人手評価はクラウドワークスを用いて実施し，計 218 名が参加した．各参加者は庭の評価もしくは対話の評価を実施し，庭の評価には計 100 名，対話の評価には計 118 名が参加した．庭の評価では，各参加者は庭の画像を基に各項目に対する回答をその理由と共に回答した．対話の評価では，各参加者は発話とブロック操作，および作成過程の動画を基に各項目に対する回答をその理由と共に回答した．

1) https://huggingface.co/rinna/japanese-gpt-neox-3.6b-instruction-ppo

表 1　1 対話あたりのアクション数（発話，ブロック設置，ブロック破壊）および庭に使用された総ブロック数．各列の最大値を**太字**，最小値を下線で示す．

エージェント	発話数	ブロック設置数	ブロック破壊数	庭のブロック数
GPT-4o	**88.7**	294.6	220.1	173.0
Gemini 2.0 Flash	43.0	641.1	515.2	246.6
DeepSeek-V3	86.3	**779.2**	**623.8**	**250.6**
Llama 3.3 70B	72.4	366.2	315.4	155.1
Qwen3 32B	49.4	469.7	406.3	161.4
人間	24.9	339.8	234.4	205.5

表 2　人手評価の平均結果．各評価指標について，（人間を除く）最も高いスコアを**太字**，2 番目に高いスコアを下線で示す．「作業中盤」は建築途中の庭，「作業終了後」は完成後の庭を指す．

エージェント	ユニークさ	美しさ	自然性	主張の積極性	合意形成度	意見反映度
GPT-4o	4.77	4.17	3.60	5.46	4.13	4.55
Gemini 2.0 Flash	4.81	**4.40**	**5.03**	**5.64**	**4.78**	**4.95**
DeepSeek-V3	**5.14**	4.16	4.22	5.50	4.65	4.72
Llama 3.3 70B	3.92	3.33	3.92	5.28	4.30	4.45
Qwen3 32B	4.50	3.94	3.98	5.03	4.33	4.80
人間（作業中盤）	4.49	3.80	-	-	-	-
人間（作業終了後）	5.15	5.08	4.84	5.03	4.83	4.51

4.2　評価指標

共同建築タスクにおける性能評価のため，客観指標と主観的な人手評価の両方を用いる．評価対象は最終的な庭に加え，建築過程での対話も含む．

客観評価では，以下の 4 指標を用いる：

発話数　対話中に生成された発話回数の合計．
ブロック設置数　対話中に設置されたブロック数．
ブロック破壊数　対話中に破壊されたブロック数．
庭のブロック数　最終的な庭に使用されたブロック総数．

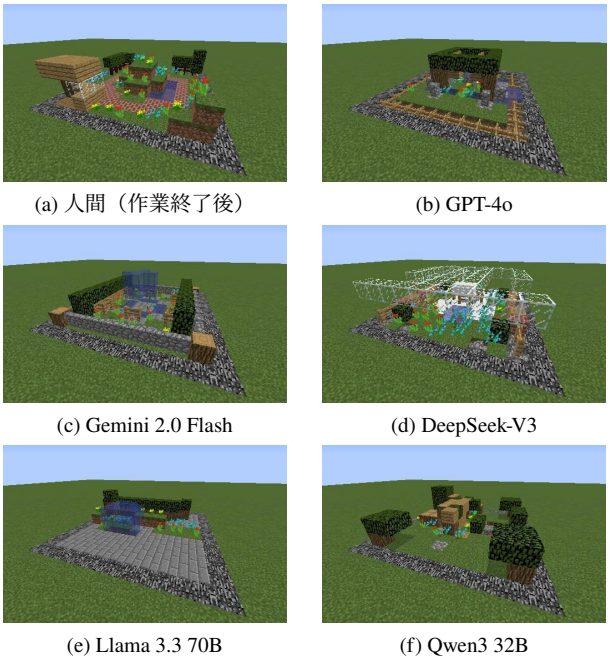
主観評価では，庭と対話それぞれに指標を設定する．庭については，「独創的で美しい庭の作成」を目的とするタスク設定 [13] を踏まえ，ユニークさと美しさの 2 指標で評価する：

ユニークさ　庭が個性的かどうか．
美しさ　庭が美しいかどうか．

対話については，共同作業の質に関わる意思疎通や相互理解を捉えるために，以下の 4 指標を用いる：

自然性　対話が自然かどうか．
主張の積極性　自身の意見を主張しているか．
合意形成度　意見のすり合わせが行われているか．
意見反映度　意見のすり合わせの結果が庭のデザインに反映されているか．

これらの主観評価は，7 段階リッカート尺度（1：最低，7：最高）で実施した．



(a) 人間（作業終了後）　(b) GPT-4o
(c) Gemini 2.0 Flash　(d) DeepSeek-V3
(e) Llama 3.3 70B　(f) Qwen3 32B

図 2　人間およびエージェントによる庭の例．

4.3　結果

表 1 は，各対話における発話数，ブロック設置・破壊数，および最終的な総ブロック数を示す．エージェントは人間より発話数が多く，特に GPT-4o と DeepSeek-V3 は人間の約 3.5 倍であった．また，Gemini 2.0 Flash と DeepSeek-V3 は設置・破壊が多く，最終的なブロック数も大きい．一方，人間は少ない発話・操作で庭を完成させており，効率的な作業が示唆される．

表 2 はクラウドソーシングによる人手評価の結果である．庭の評価では，ユニークさは DeepSeek-V3

表 3　各クラスタのラベルと，各エージェント・人間におけるクラスタ別発話割合（%）．各値は，各エージェント，人間の総発話数に対する発話の割合を示す．各列について，最大値を**太字**，2 番目に高い値を下線で示す．

ID	クラスタラベル	GPT-4o	Gemini 2.0 Flash	DeepSeek-V3	Llama 3.3 70B	Qwen3 32B	人間
1	肯定的反応	0%	0%	0%	0%	0%	**48.39%**
2	花の追加提案	**25.38%**	**24.56%**	6.55%	**27.85%**	20.34%	2.61%
3	モダンデザイン提案	20.76%	12.34%	2.90%	5.53%	6.98%	3.01%
4	石材装飾提案	5.13%	10.71%	1.51%	9.81%	2.83%	2.81%
5	素材特化デザイン	4.62%	0.35%	0.46%	3.39%	1.11%	30.52%
6	自然要素追加	11.22%	16.07%	5.21%	4.77%	11.44%	1.20%
7	最終仕上げ提案	1.58%	1.16%	**51.27%**	5.94%	3.74%	0.60%
8	休憩スペース提案	5.02%	0.93%	5.91%	0.62%	14.57%	1.20%
9	建物レイアウト提案	6.26%	3.61%	8.75%	9.47%	11.84%	4.42%
10	水の要素提案	11.84%	23.98%	7.53%	12.99%	**22.06%**	4.62%
11	池周辺装飾	8.18%	6.29%	9.91%	19.63%	5.06%	0.60%

（5.14）が最も高く，人間（5.15）に近い値を示した．美しさは Gemini 2.0 Flash（4.40）が最高で，次いで GPT-4o（4.17）であったが，人間（5.08）との差は残った．図 2a–2f に，各手法で作成された庭の例を示す．人間，GPT-4o，Gemini 2.0 Flash，DeepSeek-V3，Llama 3.3 70B，Qwen3 32B はいずれもブロックを効果的に設置している．ただし，Llama 3.3 70B と Qwen3 32B はブロック数が少なく，庭として必要な構造要素が不足しているため，特に美しさの評価が低くなった可能性がある．

対話の評価では，Gemini 2.0 Flash が自然性（5.03），主張の積極性（5.64），合意形成度（4.78），意見反映度（4.95）の全指標において，エージェントの中で最高であった．一方，合意形成度は人間を下回り，意見のすり合わせに向けた同意・不同意の明示などが十分でない可能性がある．

5　エージェント・人間の発話の比較

エージェントと人間の発話傾向の違いを発話単位で分析するため，発話のクラスタリングを行った．具体的には，エージェントと人間の各発話について SentenceBERT [2] でベクトル化し，K-means 法に基づくクラスタリングを実施した．クラスタ数は Silhouette Score と Davies-Bouldin Index に基づき 11 とした．各クラスタの解釈のため，セントロイドに最も近い発話 10 件を入力とした GPT-4o によるラベル付けを実施した．表 3 に，各クラスタのラベルとクラスタ別発話割合を示す．

表 3 より，エージェントと人間の発話分布には顕著な差異が見られる．特にクラスタ 1（肯定的反応）は人間発話のみで構成され，また人間はクラスタ 5（素材特化デザイン）の割合も高い．これらは，人間が「短い肯定」と「具体的な提案」を組み合わせて共通基盤を効率的に形成・更新している可能性を示す．一方，エージェントの発話はクラスタ 2（花の追加提案）やクラスタ 10（水の要素提案）など提案に関するクラスタを中心に構成される傾向がある．また，DeepSeek-V3 は発話の 51.27%がクラスタ 7（最終仕上げ提案）に属していた．

以上より，人間は肯定的反応と提案を組み合わせて対話する一方で，エージェントは提案を多用し，肯定的反応の活用が乏しいことが確認された．

6　おわりに

本研究では，Action-Utterance Model の行動選択フローを Creative Agents へ組み込むことでエージェントを構築し，Minecraft における共同建築タスクをシミュレーションした．生成された庭と対話の評価から，最先端 LLM を用いたエージェントが対話ベースの共同作業を遂行し，一定の創造性と品質を持つ庭を作成できることを示した．

一方，人間との比較により，エージェントは人間と同等の成果に到達するために発話回数・文字数が多く，コミュニケーション効率に課題があることが分かった．また，発話の分析から，人間は肯定的反応や提案を効果的に用いる一方で，エージェントは提案を多用し，肯定的反応の活用が乏しいことが確認された．エージェントは人間との共同作業において，提案に加えて簡潔な肯定表現を用いて，相互理解を促進することが重要である可能性がある．

今後は，対話行為やグラウンディング行為の観点から分析を深めることが重要である [19, 20, 21]．さらに，ブロック操作のために内部で生成されるコードの評価や，プロンプト設計の改良および強化学習の導入 [22, 23] による簡潔で効果的な対話・行動戦略の最適化，ターン制でなくリアルタイムに意思決定できるエージェントの構築を目指したい．

2）https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

謝辞

本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けた．

参考文献

[1] OpenAI. GPT-4 Technical Report. 2023.

[2] Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. Clarifying the Dialogue-Level Performance of GPT-3.5 and GPT-4 in Task-Oriented and Non-Task-Oriented Dialogue. In **Proceedings of the AI-HRI Symposium at AAAI-FSS 2023**, pp. 182–186, October 2023.

[3] Chen Zhang, Xinyi Dai, Yaxiong Wu, Qu Yang, Yasheng Wang, Ruiming Tang, and Yong Liu. A Survey on Multi-Turn Interaction Capabilities of Large Language Models. **arXiv preprint arXiv:2501.09959**, 2025.

[4] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. Learning Symmetric Collaborative Dialogue Agents with Dynamic Knowledge Graph Embeddings. In **Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics**, pp. 1766–1776, 2017.

[5] Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. CoDraw: Collaborative Drawing as a Testbed for Grounded Goal-driven Communication. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 6495–6513, 2019.

[6] Shuwen Qiu, Song-Chun Zhu, and Zilong Zheng. MindDial: Belief Dynamics Tracking with Theory-of-Mind Modeling for Neural Dialogue Generation. In **Proceedings of First Workshop on Theory of Mind in Communicating Agents**, 2023.

[7] Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative Dialogue in Minecraft. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 5405–5415, 2019.

[8] Haruna Ogawa, Hitoshi Nishikawa, Takenobu Tokunaga, and Hikaru Yokono. Gamification Platform for Collecting Task-oriented Dialogue Data. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 7084–7093, 2020.

[9] Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. MindCraft: Theory of Mind Modeling for Situated Dialogue in Collaborative Tasks. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 1112–1125, 2021.

[10] Chi Zhang, Penglin Cai, Yuhui Fu, Haoqi Yuan, and Zongqing Lu. Creative Agents: Empowering Agents with Imagination for Creative Tasks. **arXiv preprint arXiv:2312.02519**, 2023.

[11] Jonathan Gray, Kavya Srinet, Yacine Jernite, Haonan Yu, Zhuoyuan Chen, Demi Guo, Siddharth Goyal, C. Lawrence Zitnick, and Arthur Szlam. CraftAssist: A Framework for Dialogue-enabled Interactive Agents. **arXiv preprint arXiv:1907.08584**, 2019.

[12] Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. Learning to execute instructions in a Minecraft dialogue. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 2589–2602, 2020.

[13] Takuma Ichikawa and Ryuichiro Higashinaka. Analysis of Dialogue in Human-Human Collaboration in Minecraft. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 4051–4059, 2022.

[14] Takuma Ichikawa and Ryuichiro Higashinaka. Modeling Collaborative Dialogue in Minecraft with Action-Utterance Model. In **Proceedings of the 13th IJCNLP-AACL 2023 Student Research Workshop**, pp. 75–81, 2023.

[15] Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, et al. Interactive Grounded Language Understanding in a Collaborative Environment: IGLU 2021. In **Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track**, Vol. 176, pp. 146–161, 2022.

[16] Altera.AL. Project Sid: Many-agent simulations toward AI civilization. **arXiv preprint arXiv:2411.00114**, 2024.

[17] Justin Chiu, Wenting Zhao, Derek Chen, Saujas Vaduguru, Alexander Rush, and Daniel Fried. Symbolic Planning and Code Generation for Grounded Dialogue. In **Proceedings of the 2nd Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning**, pp. 43–53, 2023.

[18] Takuma Udagawa and Akiko Aizawa. A Natural Language Corpus of Common Grounding under Continuous and Partially-Observable Context. In **Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence**, pp. 7120–7127, 2019.

[19] David Traum. A Computational Theory of Grounding in Natural Language Conversation. **PhD thesis, University of Rochester**, 1994.

[20] Herbert H Clark. **Using language**. Cambridge university press, 1996.

[21] Biswesh Mohapatra, Seemab Hassan, Laurent Romary, and Justine Cassell. Conversational Grounding: Annotation and Analysis of Grounding Acts and Grounding Units. **arXiv preprint arXiv:2403.16609**, 2024.

[22] Bin Hu, Chenyang Zhao, Pu Zhang, Zihao Zhou, Yuanhang Yang, Zenglin Xu, and Bin Liu. Enabling Intelligent Interactions between an Agent and an LLM: A Reinforcement Learning Approach. **arXiv preprint arXiv:2306.03604**, 2023.

[23] Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Christopher Leung, Jiajie Tang, and Jiebo Luo. LLM-Rec: Personalized Recommendation via Prompting Large Language Models. **arXiv preprint arXiv:2307.15780**, 2023.