Investigating Feasibility of Large Language Model Agent Collaboration in Minecraft and Comparison with Human-Human Collaboration

Yuki Hirota **Ryuichiro Higashinaka**
Graduate School of Informatics, Nagoya University, Japan
hirota.yuki.a6@s.mail.nagoya-u.ac.jp
higashinaka@i.nagoya-u.ac.jp

Abstract

In recent years, there has been growing interest in agents that collaborate with humans on creative tasks, and research has begun to explore such collaboration within Minecraft. However, most existing studies on agents in Minecraft focus on scenarios where an agent constructs objects independently on the basis of given instructions, making it difficult to achieve joint construction through dialogue-based cooperation with humans. Prior work, such as the Action-Utterance Model, used small-scale large language models (LLMs), which resulted in limited accuracy. In this study, we attempt to build an agent capable of collaborative construction using LLMs by integrating the framework of the Action-Utterance Model with that of Creative Agents, which leverages more recent and powerful LLMs for more accurate and flexible building. We had two agents conduct the Collaborative Garden Task through simulations and evaluate both the generated gardens and the dialogue content. Through this evaluation, we confirm that the agents are capable of producing gardens with a certain level of quality and can actively offer suggestions and assert their opinions. Furthermore, we conduct a comparative analysis with human-human collaboration to identify current challenges faced by agents and to discuss future directions for improvement toward achieving more human-like cooperative behavior.

1 Introduction

In recent years, the rapid advancement of large language models (LLMs), such as GPT-4 (OpenAI, 2023), has significantly improved dialogue agents (Iizuka et al., 2023; Zhang et al., 2025). Amid this progress, research on dialogue agents designed to collaborate with humans in cooperative tasks has been actively pursued, aiming to construct more advanced dialogue agents (He et al., 2017; Kim et al., 2019; Qiu et al., 2023).

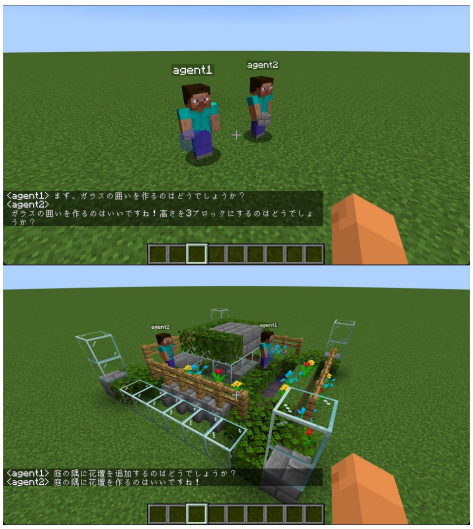


Figure 1: Visualization of agents collaboratively constructing garden. Agents are based on Gemini 2.0 Flash. Refer to Section 4.1 for how agents are constructed.

Many embodied dialogue agents designed for collaborative tasks are studied in virtual environments rather than in the real world, primarily due to cost considerations. Among these, Minecraft has been widely used as a research platform (Narayan-Chen et al., 2019; Ogawa et al., 2020; Bara et al., 2021). However, existing studies focus mainly on agents constructing structures independently (Zhang et al., 2023) or solving specific problems (Gray et al., 2019; Jayannavar et al., 2020), rather than engaging in creative, collaborative tasks. To address this gap, Ichikawa and Higashinaka (2022) proposed the Collaborative Garden Task, where agents and humans cooperatively design a garden through discussion. Their study leveraged human interaction data to fine-tune a small-scale LLM and constructed the Action-Utterance Model (Ichikawa and Higashinaka, 2023), which autonomously performs the Collaborative Garden Task. However, due to limitations in training data and insufficient modeling, this approach did not

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1333–1347

achieve satisfactory performance.

In recent years, LLMs have achieved notable progress in tasks involving the construction of predefined structures based on textual instructions. These advancements suggest that, with appropriate agent design, we may be able to enable creative and collaborative tasks with LLMs. Therefore, in this study, we build collaborative agents using state-of-the-art LLMs and investigate whether these agents can cooperatively construct a garden. Specifically, our approach applies the methodology of the Action-Utterance Model to Creative Agents (Zhang et al., 2023), a framework that leverages LLMs to construct high-quality structures from a single instruction. Figure 1 illustrates how the agents work together to collaboratively construct a garden.

In this study, we simulate dialogues between agents performing the Collaborative Garden Task and evaluate the performance of the constructed agents through subjective human assessments of the resulting gardens and dialogues. Additionally, we compare the gardens and dialogue content generated by the agents with those produced by humans performing the same task. On the basis of these results, we derive insights that can inform future improvements to the agents. Note that the target language of this study is Japanese.

The main contributions of this study are as follows:

- We built agents capable of performing a dialogue-based Collaborative Garden Task in Minecraft using state-of-the-art LLMs, and we demonstrate that these agents can collaboratively construct gardens with a certain level of quality.
- Through comparative analysis with human participants, we found that humans were able to achieve high-quality gardens and natural dialogue using short and concise utterances, whereas agents required significantly more utterances and characters to attain comparable results.
- We identified that a key difference between humans and agents lies in the extent to which they share common ground, and whether they can expand this foundation through affirmative expressions during dialogue.

Note that, following the definition of the Collaborative Garden Task by Ichikawa and Higashinaka (2022) and to allow comparisons with human-human collaboration, we focus on a setting in which two agents cooperate to create a garden. Note also that this study primarily aims to analyze the behavioral tendencies of agents in creative collaboration, rather than to propose a new method of collaboration.

2 Related Work

Related studies in this research area include work on agents that perform actions in Minecraft and those that engage in collaborative tasks using LLMs. In this section, we review these studies.

2.1 Research on Agents Performing Actions in Minecraft

Minecraft is a sandbox-style 3D game that allows players and bots to interact freely with the environment, and its multiplayer functionality enables collaborative tasks. This has led to various studies and challenges, such as the IGLU challenge (Kiseleva et al., 2022a,b), as well as multi-agent simulations (Altera.AL, 2024). Building on this, Fan et al. (2022) proposed MineCLIP, which aligns agent behavior with natural language instructions using a CLIP-style embedding space, allowing agents to learn task-relevant actions from text.

Focusing on block-based construction in Minecraft, Zhang et al. (2023) constructed Creative Agents, a framework that enables the construction of structures in Minecraft on the basis of textual instructions. When generating structures, Creative Agents utilizes an LLM to first determine the types of blocks and components needed and then generates a sequence of actions to construct the structure. This approach enables the agent to accurately interpret textual instructions and construct an entire structure from a single command. However, since Creative Agents is designed specifically for constructing structures from a single command, it lacks the ability to engage in context-aware interactions. Additionally, because of this nature, it cannot make partial modifications to existing structures or resume work from an intermediate state. In this study, we extend Creative Agents to construct an agent that can collaborate through dialogue while considering contextual information.

2.2 Research on Agents Performing Collaborative Tasks Using LLMs

Chiu et al. (2023) constructed an agent capable of solving the OneCommon task (Udagawa and

Aizawa, 2019), in which agents engage in collaborative dialogue to identify a target dot in their views. Utilizing GPT-4, the agent demonstrated a notable improvement in performance. Chaturvedi et al. (2024) constructed an agent that acts as a builder in the IGLU task (Kiseleva et al., 2022a,b). In the task, two participants take on the roles of an architect and a builder. The architect provides instructions on the next actions to be taken on the basis of a blueprint and the current state of the structure being built. The builder follows these instructions to construct the target structure. They fine-tuned an LLM using a training approach where the model learns to process instructions, actions, and corrections in a conversation.

Ichikawa and Higashinaka (2022) proposed the Collaborative Garden Task, a task in which humans or agents create a garden in Minecraft, and created an agent performing this task by fine-tuning a small-scale LLM (`japanese-gpt-neox-3.6b-instruction-ppo`[1]). They proposed the Action-Utterance Model (Ichikawa and Higashinaka, 2023), which outputs both actions and utterances on the basis of contextual information. In their approach, agents take turns, and at each turn, the model autonomously determines the appropriate type of next action on the basis of dialogue history and other contextual information. Their results indicate that the Action-Utterance Model generates more appropriate actions and utterances than a random baseline. However, their work did not leverage LLMs such as those used in Creative Agents.

3 Approach

In this study, we propose an approach that integrates two frameworks: the Action-Utterance Model and the Creative Agents framework, in order to build agents capable of performing the Collaborative Garden Task in Minecraft.

The Action-Utterance Model takes dialogue history and world state as input, selects the next action type—such as an utterance, block operation, skip, or task completion—and generates the corresponding output for that action. By repeating this process, the model enables interactive communication, allowing it to handle utterances and block operations in a sequential and integrated manner.

In contrast, Creative Agents is a non-interactive model that leverages state-of-the-art LLMs to construct high-quality structures from a single textual instruction. While it excels at selecting block types and determining placements, it does not support interaction through dialogue or step-by-step construction.

In our approach, we enhance Creative Agents and integrate it into the Action-Utterance Model framework by combining the strengths of both models through prompt extensions. Specifically, we enable Creative Agents to perform the same type of sequential control over utterances and block operations as in the Action-Utterance Model, through the following modifications:

- We introduce a procedural flow that allows the model to first select the type of next action (CHAT, BLOCK, SKIP, or FINISH) and then switch between generating utterances and performing block operations accordingly.
- We replace the single-instruction input used in conventional Creative Agents with dialogue history and world state as contextual input.
- We extend the model's capabilities to generate not only block operations but also utterances appropriate to the given context.

In collaborative tasks, it is generally assumed that each human or agent has its own intention toward the shared goal (Grosz and Sidner, 1990). However, this assumption was not incorporated into the Action-Utterance Model. To appropriately model collaboration, it is essential for each agent to hold an intention regarding its task goal. To this end, we incorporate information about the "ideal garden"—the intended goal of the task—into the agent's context. This information includes components, as well as the names of blocks to be used. By referring to this information, agents can generate utterances and perform block operations with a clear intention toward the goal. In this study, we generated the ideal garden information by prompting GPT-4o to describe images of gardens created by humans.

By combining the high-accuracy generative capabilities of Creative Agents with the sequential decision-making framework of the Action-Utterance Model, and further incorporating ideal garden information as the task goal, we build agents capable of engaging in interactive and creative collaborative construction.

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

We present the extended prompt in Appendix A. While the prompt is written in English following the prompt design of Creative Agents, we instruct the agent to produce utterances in Japanese.

4 Experiments

In this section, we describe experiments conducted to evaluate the performance of the agent created by the approach described in Section 3. We conducted the experiments after obtaining approval through the ethics review process at our affiliated institution.

4.1 Experimental Procedure

Ideally, the evaluation should involve interactions between humans and the agent. However, such an evaluation would be costly. Therefore, as an initial step, we had two agents conduct the Collaborative Garden Task through interactions, simulating the collaborative process. We then assessed the created gardens and dialogue content through subjective human evaluation via crowdsourcing.

To build the agents, we used five different LLMs as base models: GPT-4o, Gemini 2.0 Flash, DeepSeek-V3, Llama 3.3 70B Instruct, and Qwen3 32B. We selected these LLMs in consideration of a range of factors, including performance, model size, and openness. For example, GPT-4o and Gemini 2.0 Flash are high-performing commercial models, while DeepSeek-V3, Llama 3.3 70B Instruct, and Qwen3 32B are relatively large open models. By using LLMs with diverse capabilities and design philosophies, we enabled a multifaceted analysis of differences in utterance patterns and behavior across models. We conducted all dialogues in Japanese and instructed each agent to produce utterances in Japanese. See Appendix B for details on inference settings and computational resources.

For the dialogue and garden generation, we had each agent perform the Collaborative Garden Task 20 times, resulting in a total of 100 gardens and corresponding dialogue histories. We then evaluated both the generated gardens and the dialogues. Additionally, as an upper bound, we included gardens and dialogues created by human participants performing the task. Specifically, we randomly selected 20 human dialogues from the Collaborative Garden Task Corpus (Ichikawa and Higashinaka, 2022) and extracted segments from the middle and end of each dialogue[2]. We included these human-created gardens and dialogues in the evaluation alongside those generated by the agents.

4.2 Simulation Environment Construction

To evaluate the performance of the agents, we required a simulation environment where agents could interact through dialogue while collaboratively creating a garden. Ideally, each agent would control an in-game avatar in the actual Minecraft environment, moving it around and placing blocks to create the garden. However, avatars often failed to determine optimal paths to their destinations and became stuck on obstacles, causing instability in movement. To address this issue, we constructed our own virtual Minecraft environment, where only the dialogue and the world state were managed. In this virtual environment, we did not simulate avatar movement. Instead, we directly applied block operations to the world state. This approach ensured that avatars could place blocks as intended and allowed us to evaluate their dialogue and block operation abilities without external factors.

4.3 Evaluation Metrics

To evaluate the performance of agents in the Collaborative Garden Task, we used both objective behavioral metrics and subjective human judgments. We assessed not only the gardens constructed by the agents but also the dialogues exchanged during the construction process.

For the objective evaluation, we analyzed each dialogue using the following four metrics, which capture the number of actions taken and the blocks used in the resulting gardens:

Number of Utterances The total number of utterances produced during each dialogue.

Block Placements The number of blocks placed during each dialogue.

Block Destructions The number of blocks removed during each dialogue.

Blocks Used in Garden The total number of blocks used in the final garden.

In addition, to subjectively evaluate the gardens and dialogues generated by the agents, we defined separate evaluation metrics for each.

[2]Agent-generated gardens are expected to be of lower quality compared with the final gardens created by humans. To identify which stage of the human-built gardens the agent outputs correspond to in terms of quality, we also included mid-stage gardens as reference points.

Agent	Number of Utterances	Block Placements	Block Destructions	Blocks Used in Garden
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
Human	24.9	339.8	234.4	205.5

Table 1: Number of actions per dialogue and total number of blocks used in garden. Actions include utterances, block placements, and block destructions. We highlighted highest value in each column in **bold** and lowest with underlines.

Agent	Uniqueness	Beauty	Naturalness	Assertiveness	Consensus	Reflection
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
Human (Mid-task)	4.49	3.80	-	-	-	-
Human (Post-task)	5.15	5.08	4.84	5.03	4.83	4.51

Table 2: Average results of human evaluation. For each evaluation metric, we highlighted highest-scoring agent (excluding human) in **bold** and second highest with underlines. "Mid-task" refers to intermediate state of gardens during construction, while "Post-task" refers to final state of completed gardens.

For the evaluation of the gardens, we considered the objectives of the Collaborative Garden Task, which aim to create unique and aesthetically pleasing gardens (Ichikawa and Higashinaka, 2022). On the basis of these objectives, we defined two evaluation metrics:

Uniqueness The degree to which the garden is distinctive and original.

Beauty The extent to which the garden appears aesthetically pleasing.

Note that Rahimi et al. (2023) proposed more detailed subdimensions for evaluating creativity in Minecraft structures, such as Elaboration, Originality, Aesthetics, Complexity, Novel Use of Materials, and Realism. However, following the definition of the Collaborative Garden Task by Ichikawa and Higashinaka (2022), we evaluated the gardens using only the two criteria described above.

Additionally, for the evaluation of the dialogues, we considered it essential to assess the smoothness of communication and the degree of mutual understanding to measure the quality of collaboration. From this perspective, we defined four evaluation metrics:

Naturalness The degree to which the dialogue flows naturally.

Assertiveness The degree to which the agent actively expresses its opinions in the dialogue.

Consensus The degree to which opinions are negotiated and aligned within the dialogue.

Reflection The degree to which the negotiated opinions are reflected in the garden design.

For these subjective evaluations, we conducted subjective assessments (human evaluations) using a 7-point Likert scale, ranging from 1 (worst) to 7 (best). Further details of the human evaluation procedure are provided in Appendix C.

4.4 Results

Table 1 shows the number of actions per dialogue—namely, utterances, block placements, and block destructions—as well as the total number of blocks used in the final garden. The table indicates that the agents produced more utterances than the humans; in particular, GPT-4o and DeepSeek-V3 generated approximately three to four times as many utterances as the human participants. Moreover, Gemini 2.0 Flash and DeepSeek-V3 exhibited notably higher numbers of block placements and destructions compared with the other agents and humans, resulting in the largest number of blocks used in the final garden. Although the agents produced more utterances and actions than the humans, this high frequency of interaction does not necessarily indicate more effective participation, but may instead reflect verbose or redundant communication. In contrast, human participants completed their

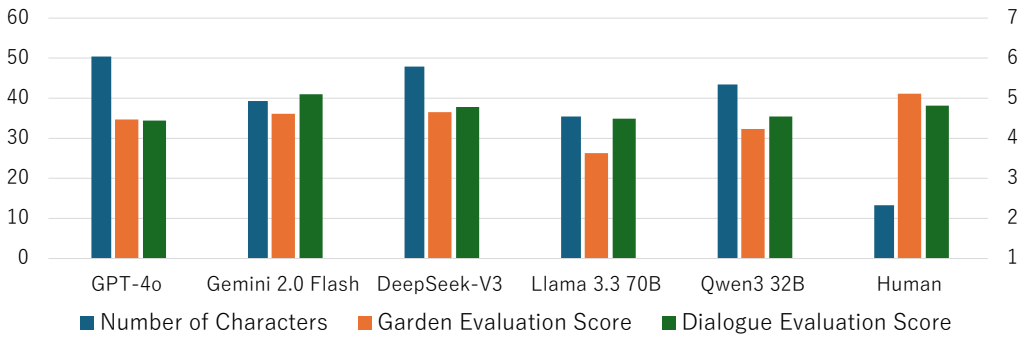


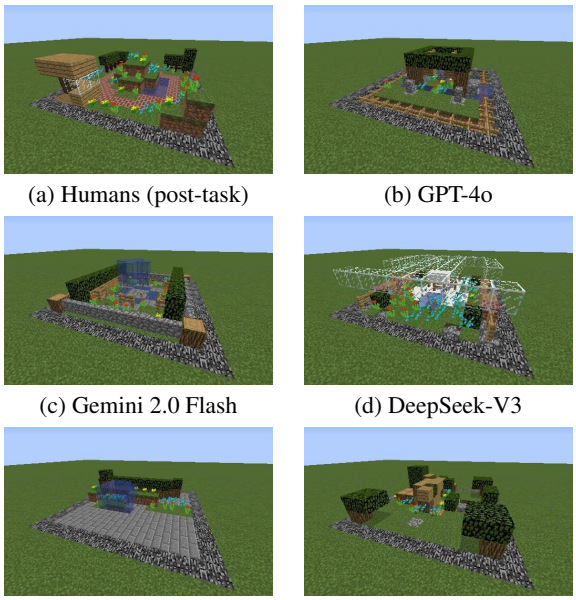
Figure 2: Average number of Japanese characters per utterance and evaluation scores for gardens and dialogues. Garden evaluation score was computed as average of Uniqueness and Beauty, while dialogue evaluation score was computed as average of Naturalness, Assertiveness, Consensus, and Reflection. Left y-axis corresponds to number of characters per utterance, while right y-axis corresponds to evaluation scores.

gardens with relatively fewer utterances and block operations, which implies more efficient coordination of dialogue and construction. This trend may be reflecting the difference in how common ground is established: whereas agents tend to explicitly confirm each proposal through utterances, humans often establish mutual understanding implicitly and proceed with minimal verbal confirmation.

Table 2 presents the results of the human evaluation by crowdsourcing, which involved a total of 218 participants, each evaluating one or more gardens and dialogues.

In the evaluation of gardens, DeepSeek-V3 achieved the highest score among the agents in terms of Uniqueness, with a score of 5.14, followed by Gemini 2.0 Flash with 4.81. Notably, DeepSeek-V3's score was nearly equivalent to the human post-task score of 5.15, suggesting that it was capable of producing gardens with a distinct character. In contrast, for Beauty, Gemini 2.0 Flash obtained the highest score among the agents at 4.40, followed by GPT-4o at 4.17. However, the human post-task score was 5.08, indicating that while the agents were able to create gardens with a certain degree of aesthetic appeal, achieving the same level of beauty as human-created gardens remains a challenge.

In the evaluation of dialogues, Gemini 2.0 Flash achieved the highest scores among the agents in all four categories: Naturalness (5.03), Assertiveness (5.64), Consensus (4.78), and Reflection (4.95). Notably, its scores for Naturalness, Assertiveness, and Reflection exceeded those of the human participants. These results suggest that the agent actively expressed its own ideas and proposals, facilitating lively exchanges of opinions compared with the humans. However, the score for Consensus was lower than that of the human participants, indicating that despite its assertiveness, the agent may not have sufficiently responded to the partner's utterances with clear agreement or disagreement.



(a) Humans (post-task) (b) GPT-4o

(c) Gemini 2.0 Flash (d) DeepSeek-V3

(e) Llama 3.3 70B (f) Qwen3 32B

Figure 3: Examples of gardens created by humans and agents.

Figure 2 shows the average number of Japanese characters per utterance along with the evaluation scores for the gardens and dialogues. From the figure, we observe that the agents tended to produce more utterances and longer text per utterance compared with the humans. This result suggests that the agents were more likely to generate verbose utterances and may have been less efficient than the human participants. In contrast, the humans progressed through the dialogue with signif-

A1: Should we start by focusing on a specific part of the garden, or should we begin by selecting materials?
B1: To plan the garden, shall we start with the pathways or the basic structure? Which do you prefer?
A2: How about laying a cobblestone path as the foundation of the garden structure?
B2: Shall we start by creating a cobblestone path?
A3: [Places cobblestone]
B3: Next, how about placing decorative hedges around the garden?

Figure 4: Example dialogue from GPT-4o. A and B represent speakers, and numbers indicate serial numbers for utterances and actions. Actions enclosed in square brackets represent block operations. This dialogue was originally in Japanese and has been translated into English by authors.

A1: How about starting by building a glass enclosure?
B1: Building a glass enclosure sounds good! How about making it 3 blocks tall?
A2: [Places glass]
B2: [Places dirt]
A3: What do you think about turning the glass enclosure into a greenhouse-like environment?
B3: [Places glass]

Figure 5: Example dialogue from Gemini 2.0 Flash. Refer to Figure 4 for notations.

icantly fewer utterances while achieving equal or better evaluation scores for both the gardens and dialogues. This finding suggests that the humans efficiently communicated within a limited number of exchanges and achieved high performance with less verbal input. These results imply that the humans effectively communicated on the basis of a shared understanding; it is important for agents to also establish and utilize common ground, enabling them to convey intentions even with fewer utterances.

4.5 Case Study

Figures 3(a)–3(f) present examples of gardens created using each method. The gardens created by the humans (Figure 3(a)), GPT-4o (Figure 3(b)), Gemini 2.0 Flash (Figure 3(c)), DeepSeek-V3 (Figure 3(d)), Llama 3.3 70B (Figure 3(e)), and Qwen3 32B (Figure 3(f)) all exhibited effective placement of individual elements. However, the gardens created by Llama 3.3 70B (Figure 3(e)) and Qwen3 32B (Figure 3(f)) contained relatively few blocks and lacked structural elements typically expected in a garden. We believe this contributed to their comparatively lower evaluation scores.

We present an example dialogue from GPT-4o

A1: Is there anything you would like to build?
B1: We haven't built anything in a Japanese style for a while, so I'd like to try that.
A2: How about a Japanese garden?
B2: That sounds great!
A3: A Japanese garden usually has a pond, right?
B3: Exactly!

Figure 6: Example dialogue from humans. Refer to Figure 4 for notations.

in Figure 4, Gemini 2.0 Flash in Figure 5, and from the human participants in Figure 6. In the dialogue from GPT-4o (Figure 4), B2 proposed, "Shall we start by creating a cobblestone path?" However, in A3, the agent did not respond to this proposal and instead proceeded to place a block. Additionally, while most utterances took the form of proposals, there were no corresponding responses to these suggestions. This resulted in a series of unilateral proposals, making the dialogue appear unnatural. In the dialogue from Gemini 2.0 Flash (Figure 5), unlike GPT-4o, the agent in B1 responds to A1 with an utterance such as "Building a glass enclosure sounds good!" However, in response to the suggestion in B1—"How about making it 3 blocks tall?"—we would expect a reply in A2. Instead, A2 immediately proceeds with an action such as "[Places glass]," despite the proposal not yet being agreed upon. This lack of response to suggestions and assertions in Gemini 2.0 Flash likely contributed to its lower Consensus score in dialogue evaluation. In contrast, in the human dialogue (Figure 6), participants explicitly expressed agreement with each other's suggestions and maintained smooth interactions with an appropriate degree of empathy. We also observed that they communicated efficiently using relatively short utterances such as "That sounds great!" or "Exactly!" These observations suggest that while the constructed agents can make proposals and assertions, they fall short in engaging in reciprocal dialogue—specifically, in building and maintaining common ground with their partners as the conversation progresses.

5 Comparing Agent and Human Utterances through Clustering

To analyze differences between the agents (GPT-4o, Gemini 2.0 Flash, DeepSeek-V3, Llama 3.3 70B, Qwen3 32B) and humans at the utterance level, we conducted clustering on individual utterances. Specifically, we used a Japanese-compatible ver-

ID	Cluster Label	GPT-4o	Gemini 2.0 Flash	DeepSeek-V3	Llama 3.3 70B	Qwen3 32B	Human
1	Affirmative response	0%	0%	0%	0%	0%	**48.39%**
2	Suggestions to add flowers	**25.38%**	**24.56%**	6.55%	**27.85%**	20.34%	2.61%
3	Suggestions for modern designs	20.76%	12.34%	2.90%	5.53%	6.98%	3.01%
4	Suggestions for stone decorations	5.13%	10.71%	1.51%	9.81%	2.83%	2.81%
5	Material-focused design suggestions	4.62%	0.35%	0.46%	3.39%	1.11%	30.52%
6	Additions of natural elements	11.22%	16.07%	5.21%	4.77%	11.44%	1.20%
7	Final touch suggestions	1.58%	1.16%	**51.27%**	5.94%	3.74%	0.60%
8	Suggestions for rest areas	5.02%	0.93%	5.91%	0.62%	14.57%	1.20%
9	Suggestions for building layout	6.26%	3.61%	8.75%	9.47%	11.84%	4.42%
10	Suggestions for water features	11.84%	23.98%	7.53%	12.99%	**22.06%**	4.62%
11	Decorations around the pond	8.18%	6.29%	9.91%	19.63%	5.06%	0.60%

Table 3: Cluster labels and proportion of utterances (%) belonging to each cluster for each agent and human participant. Each value indicates percentage of utterances within given cluster relative to total number of utterances by corresponding agent or human group. For each agent or human, we use **bold** to indicate cluster with highest proportion and underline for second highest.

sion of SentenceBERT[3] to embed each utterance—including those from agents and humans—into vector representations, and we applied the K-means algorithm for clustering[4]. By clustering utterances from agents and humans together, we enabled a direct comparison of how their utterances are distributed across clusters in terms of semantic differences. We determined the optimal number of clusters to be 11 using the Silhouette Score and the Davies-Bouldin Index.

To qualitatively interpret the contents of each cluster, we extracted the 10 utterances closest to the centroid of each cluster and used GPT-4o to label the semantic meaning of each cluster. We present in Table 3 the cluster IDs, their corresponding labels, and the proportion of utterances across clusters for each agent and human participant.

From Table 3, we observe distinct utterance tendencies between the agents and humans. In particular, Cluster 1, labeled "Affirmative response," consists entirely of utterances from the human participants. Additionally, 51.27% of all utterances from DeepSeek-V3 fall into Cluster 6, labeled "Final touch suggestions," indicating that the agent repeatedly proposed finishing touches throughout the task. We also observe that a large proportion of human utterances fall into Clusters 1 and 5. Utterances in Cluster 1 include expressions of approval toward the partner's suggestions or work, such as "I think that looks great!" Meanwhile, utterances in Cluster 5 tend to be more complex and specific, such as "How about building a modern pavilion with quartz blocks and adding a gravel path? I'd love to hear your thoughts." These examples suggest that the humans often engaged in both affirming contributions and proposing detailed design ideas.

These findings indicate that agents and humans exhibit different tendencies in their utterances. The human participants frequently used utterances such as affirmations toward the partner and concrete proposals about layout and materials. The frequent use of affirmative expressions by the humans is also evident in the human dialogue example shown in Figure 6, and it appears to play an important role in efficiently establishing common ground. In contrast, the agents did not effectively use concise and clear expressions of affirmation or empathy, which may have contributed to the more verbose dialogues and the difficulty in building common ground, as seen in Figures 4 and 5. Therefore, to enable agents to engage in more human-like dialogue, it may be beneficial for them not only to convey proposals and information, but also to use concise and effective affirmations and empathetic expressions to facilitate mutual understanding, especially when cooperating with humans. While it is not necessarily the case that behaving like humans is always the optimal strategy for agents, such behavior should be important for establishing smooth collaboration with human partners.

6 Conclusion

In this study, we aimed to construct an agent capable of creating high-quality gardens by incorporating the approach of the Action-Utterance Model into Creative Agents, and we conducted simulations of the Collaborative Garden Task. We then evaluated and analyzed both the dialogues generated during the simulations and the gardens created by the agents.

[3] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

[4] The random_state parameter was set to 42.

Our study is the first to demonstrate that agents powered by state-of-the-art LLMs can successfully engage in a dialogue-based collaborative construction task in Minecraft, producing gardens with a certain level of creativity and quality. Through comparative analysis with human participants, we found that while agents can complete the task, they require significantly more utterances and characters to reach similar outcomes, indicating inefficiencies in communication. Furthermore, our utterance-level analysis revealed that unlike humans, agents rarely use brief affirmations or empathetic expressions—an important factor in establishing and expanding common ground—highlighting a key area for future improvement in human-agent collaboration.

As future work, a more detailed analysis of dialogue content will be important, including labeling from the perspectives of dialogue acts and grounding acts (Traum, 1994; Clark, 1996; Mohapatra et al., 2024). In this study, the agents internally generate code to determine their block operations. Therefore, a quantitative evaluation of the accuracy and reliability of the code generated by the agents is necessary. In future studies, we plan to introduce measures such as code complexity and analyze the causes of code regeneration (Austin et al., 2021). In addition, to achieve more human-like collaboration, we believe it will be effective to optimize the agent's dialogue and action strategies by refining prompt design on the basis of the insights gained in this study and by incorporating reinforcement learning techniques (Hu et al., 2023; Lyu et al., 2023), such as prompt tuning to encourage more concise and efficient utterances. Furthermore, in this study, the agent operated in a turn-based manner. However, human players typically act in real time without strict turn-taking. Therefore, as future work, we aim to construct an agent capable of making real-time decisions.

7 Limitations

In this study, we limited our investigation of collaborative behavior to the Collaborative Garden Task. As a result, further work is needed to determine the extent to which our findings generalize to other types of structures, to creative and cooperative tasks in environments beyond Minecraft, and to real-world scenarios involving robots. Additionally, we based our evaluation on simulations of collaboration between agents and did not include joint tasks involving both humans and agents.

Moreover, our dialogue evaluations relied on subjective human judgments collected through crowdsourcing, which may be influenced by individual differences and potential biases among evaluators. Finally, all dialogues in this study were conducted in Japanese. It remains an open question whether agents would produce similarly verbose utterances in other languages or cultural contexts and how such behavior would be evaluated.

8 Ethical Considerations

This study includes human participant evaluations of both gardens and dialogues. Before conducting the human evaluations, we provided participants with a thorough explanation of the research purpose, evaluation procedure, handling of personal information, and confidentiality obligations. We obtained their informed consent prior to participation. In setting compensation for the evaluation tasks, we took into account the minimum wage standards in Japan and made efforts to ensure fair hourly pay. Regarding the agent's behavior and generated utterances, there is a possibility that unintended biases or inappropriate expressions may arise. To mitigate these risks, it is important to conduct sufficient pre-testing and apply filtering where necessary. Furthermore, if we apply the findings of this research to real-world systems such as robots, we must carefully examine their safety and reliability and thoroughly consider their potential impact on users. Regarding the agents evaluated in this study, we used the following large language models: GPT-4o, Gemini 2.0 Flash, DeepSeek-V3, Llama 3.3 70B Instruct, and Qwen3 32B. All these models were utilized in accordance with their respective terms of use and licenses.

9 Acknowledgments

This work was supported by JST Moonshot R&D Grant Number JPMJMS2011.

References

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

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and 1 others. 2021. Program Synthesis with Large Language Models. *arXiv preprint arXiv:2108.07732*.

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

Akshay Chaturvedi, Kate Thompson, and Nicholas Asher. 2024. Nebula: A discourse aware Minecraft Builder. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6431–6443.

Justin Chiu, Wenting Zhao, Derek Chen, Saujas Vaduguru, Alexander Rush, and Daniel Fried. 2023. 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*, pages 43–53.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. In *Proceedings of Neural Information Processing Systems*, volume 35, pages 18343–18362.

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

Barbara J. Grosz and Candace L. Sidner. 1990. Plans for Discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 417–444. MIT Press.

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

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

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

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

Shinya Iizuka, Shota Mochizuki, Atsumoto Ohashi, Sanae Yamashita, Ao Guo, and Ryuichiro Higashinaka. 2023. 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*, pages 182–186.

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

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

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, and 1 others. 2022a. Interactive Grounded Language Understanding in a Collaborative Environment: IGLU 2021. In *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, volume 176, pages 146–161.

Julia Kiseleva, Alexey Skrynnik, Artem Zholus, Shrestha Mohanty, Negar Arabzadeh, Marc-Alexandre Côté, Mohammad Aliannejadi, Milagro Teruel, Ziming Li, Mikhail Burtsev, and 1 others. 2022b. Interactive Grounded Language Understanding in a Collaborative Environment: Retrospective on Iglu 2022 Competition. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220, pages 204–216.

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

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

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

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

OpenAI. 2023. GPT-4 Technical Report.

Shuwen Qiu, Song-Chun Zhu, and Zilong Zheng. 2023. 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*.

Seyedahmad Rahimi, Justice Walker, Lin Lipsmeyer, and Jinnie Shin. 2023. Toward Defining and Assessing Creativity in Sandbox Games. *Creativity Research Journal*, 36:1–19.

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

Takuma Udagawa and Akiko Aizawa. 2019. 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*, pages 7120–7127.

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

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

A Agent Prompt

We present the prompt used to construct the agent in Listing 1. In this study, we extended the prompt design of Creative Agents to include dialogue history, world state, and ideal garden information as contextual input. We also instructed the agent to explicitly select the next action type—utterance, block operation, skip, or complete—so that it could appropriately distinguish between generating utterances and performing block operations.

Listing 1: Agent Prompt

```
You are a creative and collaborative agent(system) to build a garden in Minecraft with a partner. The task is turn-based, and you
and the partner take turns performing actions to create the ideal garden collaboratively. You are given the following information
to build a garden:
1. Action History: A record of all previous interactions, including past system and partner messages and any previous building
actions.
2. World State: The current block informations of the garden in Minecraft.
3. Ideal Garden Information: A description of the ultimate garden design you aim to achieve.

Here is the current information:
<action_history> {action_history} </action_history>
<world_state> {world_state} </world_state>
<ideal_garden_info> {ideal_garden_info} </ideal_garden_info>

Based on the current inputs and conditions, please consider the following step by step:
1. Based on the current garden state, what are the next steps to build the garden collaboratively with the partner? You need to
consider your next step from the following actions: [CHAT, BLOCK, SKIP, FINISH].
1.1. CHAT: Speak to the partner, e.g., by asking a question or giving a suggestion. You should chat moderately before placing or
breaking blocks, but avoid repeating similar questions to the partner and ensure a balance between communication and block
operations, engaging in block placement or removal at appropriate intervals.
1.2. BLOCK: Place or remove a block.
1.3. SKIP: Take no action, allowing the partner to take their turn.
1.4. FINISH: Finish the garden construction.
2. If you chose CHAT, provide a clear and concise message that reflects your understanding of the garden's layout and design. You
should provide thoughtful suggestions or ideas about changes to the garden. You need to chat in Japanese.
3. If you chose BLOCK, consider the next block to place or remove based on the current garden state and the ideal garden
information. What blocks should be placed or removed to bring the garden closer to the ideal?
4. If you chose CHAT or BLOCK, after deciding on the next step, generate Mineflayer JavaScript code to control the bot in
Minecraft.
5. If you chose SKIP, output only `<|SKIP|>`.
6. If you chose FINISH, output only `<|FINISH|>`.

// These are Mineflayer functions you can use:
bot.blockAt(position); // Return the block at `position`. `position` is `Vec3`. The block has two member variables: `name`(string
) and `position`(Vec3).
bot.chat("message"); // Send a chat message.
placeItem(bot, name, position); // Place a block at the specified position.
breakBlock(bot, position); // Break a block at the specified position.

You should then respond to me with
Explain: Explain the steps you have taken to build the garden and the reasons behind your decisions.
Plan: How to complete the task.
Code:
1) Write an async function taking the bot as the only argument.
2) To break or place blocks, or chat with the partner, use the following functions:
- Use `await breakBlock(bot, position)` to dig blocks. Do not use `bot.dig` directly.
- Use `await placeItem(bot, name, position)` to place blocks. Do not use `bot.placeBlock` directly.
- Use `bot.chat` to chat with the partner. You should chat in Japanese.
3) All coordinates of blocks are relative, You must write `const position = getBotPosition(bot);` in the first line of the
function to get absolute coordinates.
4) Anything defined outside a function will be ignored, define all your variables inside your functions.
5) Do not write infinite loops or recursive functions.
6) Do not use `bot.on` or `bot.once` to register event listeners. You definitely do not need them.
7) Name your function in a meaningful way (can infer the task from the name).
8) Do not use any other APIs not mentioned above.

You should only respond in the format as described below:
RESPONSE FORMAT:
Explain: ...
Plan:
1) ...
2) ...
3) ...
...
Code:
```javascript
// helper functions (only if needed, try to avoid them)
...
// main function after the helper functions
async function buildGarden(bot) {{
// ...
}}
```

Here is an example of java script code:
```

"BLOCK" Code Example:
```javascript
// helper function to build a house
async function buildPond(bot, position, size, blockName) {{
 for (let y = 0; y < size; y++) {{
 for (let x = 0; x < size; x++) {{
 for (let z = 0; z < size; z++) {{
 const targetPosition = position.offset(x, y, z);
 await placeItem(bot, blockName, targetPosition);
 }}
 }}
 }}
}}

// main function
async function buildGarden(bot) {{
 const position = getBotPosition(bot);
 const size = 5; // size of the house
 const blockName = 'blue_stained_glass';
 await buildGarden(bot, position, size, blockName);
}}
```

"CHAT" Code Example:
```javascript
// main function
async function chatAboutPond(bot) {{
 bot.chat("How about planting trees around the pond?");
}}
```


Please note that:
1) Never check whether you have enough blocks in inventory. I will garantee that you will be given enough blocks.
2) Always use ```const position = getBotPosition(bot);```.
3) Never define ```placeItem(bot, blockName, targetPosition)``` by yourself. We already provide a defined function.
4) In terms of the size of the garden, the kind of blocks of your selection and other details, please refer to your answers to those questions above.
5) The relative coordinates must fall within the ranges:
 - x: [0, 9]
 - y: [0, 3]
 - z: [0, 9]

Here are the names of the blocks that you can choose from:
['blue_stained_glass','brick_wall', 'stone_brick_wall', 'oak_fence', 'oak_planks', 'oak_log', 'oak_leaves', 'dandelion', 'blue_orchid', 'red_tulip', 'grass_block', 'dirt', 'cobblestone', 'bricks', 'stone_bricks', 'quartz_block', 'glass']
You CANNOT use any other blocks not listed above. Be careful the spell of blocks, you should not misspell.
LIMIT: We need to build a garden using `blue_stained_glass` instead of `water`. `blue_stained_glass` is not fluid unlike water, so avoid leaving unnecessary space around the same height.
`dandelion`, `blue_orchid`, `red_tulip` are flowers, can only place on `dirt` or `grass_block`. Before placing the flower, place a grass_block underneath it.
You should not misspell them in your code.

You should write your code within maximum length of tokens.

B Large Language Models and Computational Environment

In this study, we used five LLMs: GPT-4o (`gpt-4o-2024-08-06`), Gemini 2.0 Flash, DeepSeek-V3, Llama 3.3 70B Instruct, and Qwen3 32B. We accessed GPT-4o, Gemini 2.0 Flash, and DeepSeek-V3 via APIs, while we ran Llama 3.3 70B Instruct and Qwen3 32B in a local environment. For local execution, we used four NVIDIA RTX A6000 GPUs. Running the simulations required approximately 50 hours for Llama 3.3 70B and 30 hours for Qwen3 32B.

For inference with each model, we used the generation hyperparameters listed in Table 4.

LLM	temperature	top_p	top_k
GPT-4o	1.0	1.0	–
Gemini 2.0 Flash	1.0	0.95	64
DeepSeek-V3	1.0	1.0	–
Llama 3.3 70B	0.7	1.0	50
Qwen3 32B	1.0	1.0	50

Table 4: Hyperparameters used for inference with each model.

C Details of Human Evaluation

We conducted human evaluations of the gardens and dialogues using the CrowdWorks platform[5]. Each participant evaluated either a garden or a dialogue. Since the evaluation included free-text fields for participants to describe the basis of their ratings, we obtained consent from all participants regarding the following: "Agreement on Handling of Personal Information," "Waiver of Copyright," "Agreement on Data Usage," and "Confidentiality Agreement."

C.1 Details of Garden Evaluation

Figure 7 shows an example of a garden image used in the garden evaluation.

For the garden evaluation, we provided the following instructions. We set the expected completion time to approximately 30 minutes and the compensation to 600 yen. The instructions below are the English version translated by the authors.

```
[Overview]
For research and development purposes, we would like to
compare evaluations of structures created by a system with
those created by humans. This task will be conducted in a
structured format.

[Task Details]
- You will be shown images of gardens created in Minecraft.
- For each garden, you will be asked to evaluate whether the
garden is unique and whether it is beautiful.
```

[5] https://crowdworks.jp/

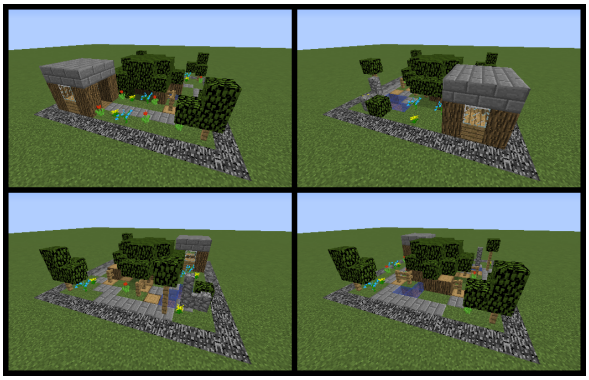


Figure 7: Example image of garden used in evaluation.

```
- First, please rate your level of agreement with the
statement: "This garden is unique." using a 7-point scale.
- Then, briefly explain the reason for your rating in
approximately 10 to 50 characters.
- Similarly, rate your agreement with the statement: "This
garden is beautiful." using a 7-point scale, and provide a
reason in about 10 to 50 characters.
- You will be asked to evaluate a total of 12 garden images.

[Important Notes]
- The estimated time required for the task is approximately 30
 minutes (this may vary between individuals).
- Responses will be collected via Google Forms. No Google
account is required to complete the form.
- To complete the task, you will need a confirmation code that
 will be displayed on the Google Form after submitting all
your responses.
- If your answers appear to ignore the questions, or if the
task is completed in an unusually short time, your submission
may be rejected. Thank you for your understanding.

[Steps to Complete the Task]
1. Click the "Start Task" button.
2. A Google Forms URL will be displayed - click the link to
access the form.
3. View each garden image on the form and respond to the
questions. (No Google account is required.)
4. After answering all questions, submit the form. A
confirmation code will appear on the screen - be sure to take
note of it.
5. Enter the confirmation code on the CrowdWorks page and
click the "Complete Task" button.
```

Below, we present examples of the evaluation criteria used in the assessment. The examples shown have been translated into English by the authors.

```
[Statement]
This garden is unique.

[Evaluation Criteria]
- Strongly disagree
- Disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Agree
- Strongly agree

[Reason]
Please provide the reason for your evaluation (10-50
characters).
```

C.2 Details of Dialogue Evaluation

For the dialogue evaluation, we provided the following instructions. We set the expected completion time to approximately 20 minutes and the compensation to 400 yen. The instructions below are

the English version translated by the authors.

```
[Overview]
For research and development purposes, we would like to
compare evaluations of system-to-system dialogues with those
of human collaborative tasks. This task will be conducted in a
 structured format.

[Task Details]
- Two individuals, referred to as A and B, are collaboratively
 creating a garden in Minecraft through chat-based dialogue
and block operations (placing and breaking blocks).
- This collaborative garden-building activity will be referred
 to as a task below.
- For each task, you will be shown the dialogue, block
operations, and an animation of the garden creation process,
both at the beginning and the end of the task.
- Based on these, please indicate your level of agreement with
 the following five statements:
 1. "The task (dialogue and block operations) between A and B
 feels natural."
 2. "A is able to express their own opinions."
 3. "B is able to express their own opinions."
 4. "A and B are negotiating their ideas for what they want to
  build together."
 5. "The result of A and B's negotiation is reflected in the
 garden design."
- For each statement, rate your agreement on a 7-point scale
and briefly explain your reasoning in approximately 10 to 50
characters.
- You will evaluate two tasks in total. Note that A and B may
not be the same individuals in both tasks - the names are
assigned for each task and should be interpreted as such.

[Important Notes]
- The estimated time required for the task is approximately 20
 minutes (this may vary between individuals).
- Responses will be collected via Google Forms. No Google
account is required to complete the form.
- To complete the task, you will need a confirmation code that
 will be displayed on the Google Form after submitting all
your responses.
- If your answers appear to ignore the questions, or if the
task is completed in an unusually short time, your submission
may be rejected. Thank you for your understanding.

[Steps to Complete the Task]
1. Click the "Start Task" button.
2. A Google Forms URL will be displayed - click the link to
access the form.
3. View each garden image on the form and respond to the
questions. (No Google account is required.)
4. After answering all questions, submit the form. A
confirmation code will appear on the screen - be sure to take
note of it.
5. Enter the confirmation code on the CrowdWorks page and
click the "Complete Task" button.
```

Below, we present a portion of the dialogue content used in the evaluation, along with examples of the evaluation criteria. While 50 actions were shown during the actual evaluation, we display only 10 actions here for brevity. The dialogue content and evaluation criteria shown below have been translated into English by the authors.

```
[Excerpt from the Dialogue]
B: Is there anything you'd like to make?
A: I haven't made anything Japanese-style in a while, so
something in that style would be nice.
B: Then how about something like a Japanese garden?
A: Let's do that!
B: A Japanese garden needs a pond, right?
A: Right!
 [Destroys block]
B: [Destroys block]
A: [Destroys block]
B: [Places grass block]

[Statement]
The task (dialogue and block operations) between A and B feels
 natural.

[Evaluation Criteria]
- Strongly disagree
- Disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Agree
- Strongly agree

[Reason]
Please provide the reason for your evaluation (10-50
characters).
```