# Benchmarking AI Agents' Creativity in Dynamic Virtual Worlds

A Technical Report
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

**Yifan Zhou**

November 20, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Yifan Zhou

Technical advisor: Yen-Ling Kuo, Department of Computer Science

# Benchmarking AI Agents' Creativity in Dynamic Virtual Worlds

Celine Zhou
University of Virginia
qmj3vs@virginia.edu

## ABSTRACT

This study benchmarks the creative capabilities of AI agents powered by large language models (LLMs), including GPT-4o, Claude, Gemini, and Llama, within the dynamic virtual environment of Minecraft. Using tasks such as house building, garden design, and decoration, the agents were evaluated based on originality, appropriateness, and aesthetic appeal through a comprehensive evaluation matrix design based on frameworks like the Torrance Tests of Creative Thinking (TTCT) and the Consensual Assessment Technique (CAT). The findings reveal distinct patterns in the creativity of LLM-powered agents, highlighting their strengths in appropriateness and sentimental matching, as well as limitations in originality and contextual adaptability. These insights provide valuable guidance for optimizing AI performance and expanding its applications in education, gaming, and creative industries.

## Introduction

Generative AI has emerged as a transformative force in technology, pushing the boundaries of what machines can achieve in creative domains. Models such as ChatGPT and DALL-E have demonstrated groundbreaking capabilities in generating text and images, fueling debates about AI's role in creativity. ChatGPT excels in language-driven tasks, from crafting stories to composing poetry. At the same time, DALL-E has set new standards for AI-driven artistry by creating intricate visual content from textual prompts. These advancements have inspired interest in the potential of AI to extend beyond single-medium creativity to integrated, multimodal applications.

Virtual worlds like Minecraft provide an ideal testing ground for exploring AI's creative problem-solving abilities. Minecraft functions like a more flexible version of LEGO, offering an open-ended platform where users can construct anything they imagine using over 800 placeable unique blocks. This vast range of components enables the creation of intricate structures and even complex mechanical devices, encouraging players to push the limits of their creativity. Minecraft's open-ended, interactive tasks mimic real-world challenges, allowing AI agents to engage in activities such as constructing architecture, gathering and managing resources, and decorating spaces to meet specific themes. Minecraft fosters limitless imagination as an open platform, enabling AI agents to experiment with creative building and mechanical design. These features make Minecraft an unparalleled environment for testing AI creativity in dynamic, interactive, and highly customizable scenarios.

By adopting interdisciplinary methodologies, this research examines the creative potential of AI through the lens of cognitive psychology, utilizing established frameworks such as the Torrance Tests of Creative Thinking (TTCT) and the Consensual Assessment Technique (CAT). The study evaluates the performance of AI agents across standardized tasks, focusing on three metrics: Originality, Appropriateness, and Aesthetic Appeal for decoration tasks. Ultimately, this research aims to provide a comparative framework for assessing AI creativity, offering insights into the strengths and limitations of current generative models in virtual environments.

## Related Work

**Cognitive Psychology and AI Creativity:** Creativity assessment frameworks like the Torrance Tests of Creative Thinking (TTCT; Torrance, 1966) [7] and the Consensual Assessment Technique (CAT; Amabile, 1982) [2] provide systematic methods for evaluating originality, usefulness, and complexity in human outputs. As shown by Chen et al. (2024) [1], these benchmarks have been extended to AI, which successfully adapted TTCT to evaluate AI-generated outputs. Such methodologies are the foundation for this study's approach to assessing creativity in virtual environments.

**AI in Virtual Worlds:** Minecraft has become an effective platform for testing AI creativity due to its open-ended, dynamic nature. Nottingham et al.(2023) [5] demonstrated how LLM-guided exploration improves reinforcement learning efficiency by hypothesizing and verifying task subgoals. Similarly, Fan et al. (2022) [3] introduced a framework that integrates a simulation suite with internet-scale knowledge to enable agents to perform diverse, language-guided tasks, showcasing AI's potential for generalist creativity. These studies highlight the flexibility of Minecraft for benchmarking AI adaptability and problem-solving.

**Mindcraft and Related Tools:** This creativity project builds on **Mindcraft** (Nottingham, 2024) [4], a project based on **Mineflayer** (PrismarineJS, n.d.) [6], which facilitates the integration of AI agents into Minecraft. The name "Mindcraft" signifies equipping agents in Minecraft with an AI "mind," enabling autonomous and intelligent actions. Mindcraft allows researchers to test AI models in tasks like crafting, building, and navigation using

standardized APIs. Projects like **Voyager** (Wang et al., 2022) [8], built on the previously mentioned framework Fan et al. (2022) [3] , further illustrate how language models can autonomously explore and adapt in virtual worlds, validating the practical applications of these tools in AI creativity research.

## AI Agent Setup


Fig. 1 AI Agent Andy

The AI agent was deployed in the Minecraft environment using the Mindcraft framework, which integrates large language models (LLMs) such as GPT-4o, Gemini, Claude, and Llama through API calls. These LLMs enable the agent to interpret and respond to user commands effectively. Minecraft's interactivity is facilitated by the Mineflayer library, which allows the agent to perform tasks such as navigating, gathering resources, and constructing structures. This combination of tools provides a robust foundation for enabling AI behavior in a dynamic and creative virtual environment.

The agent's configuration is defined in a core file named andy.json, which specifies key parameters such as the agent's name ("name": "andy"), model (e.g., "gpt-4o-mini"), and connection details for the Minecraft server, including the localhost address ("host": "localhost") and port number ("port": 55916). These parameters initialize the agent and establish its connection to the Minecraft server, allowing real-time interaction. The andy.json file also includes a conversing field, which defines behavioral guidelines for the agent, such as responding concisely, avoiding unnecessary apologies, and immediately executing user commands. These rules ensure that the agent interacts naturally and behaves as a typical Minecraft player would.

The agent's functionality is managed by several key files within the Mindcraft codebase. The agent.js file defines the Agent class, which oversees the lifecycle of the agent, including updating its state via the update method and safely terminating its processes through cleanKill. The prompter.js file handles the generation and processing of user prompts, ensuring that the agent interprets and responds to commands accurately. Additionally, the history.js file manages the agent's conversation history, retaining context across interactions through the add method and summarizing past messages to maintain efficient memory usage. The coder.js file enables the agent to execute tasks programmatically by generating and running JavaScript commands using functions such as stageCode and execute. These components collectively allow the agent to perform complex actions like moving to a specific location, collecting blocks, and placing structures in the Minecraft environment.

The integration of these tools and functions enables the seamless operation of the AI agent within Minecraft. Through the Mindcraft framework, the agent connects to a Minecraft server hosted on localhost with port 55916, allowing it to collaborate with human users on creative tasks such as building and resource management.

## Methodology

This study's methodology is structured into six key stages, as the flowchart shows. First, established psychological frameworks were used to design tasks and evaluation criteria, including the Torrance Tests of Creative Thinking (TTCT) and the Consensual Assessment Technique (CAT). Based on these, three tasks were selected: building a house, decorating a house, and designing a garden. Three prompt types—basic, instructive, and chain-of-thought—were developed to prompt creativity in AI responses. These prompts were tested across four AI models (Llama 3, GPT-4o, Claude, and Gemini) through API connections in a controlled environment. The outputs were evaluated on originality, appropriateness, and aesthetic appeal and then ranked by human participants to assess each model's creative performance.

The Torrance Tests of Creative Thinking (TTCT) and the Consensual Assessment Technique (CAT) were selected for their complementary methodologies in assessing creativity, particularly in contexts involving artificial intelligence. TTCT is a widely recognized and classical framework for measuring creative potential and divergent thinking, often applied to general populations such as students and professionals. Its structured evaluation metrics, including fluency, originality, elaboration, and flexibility, provide a standardized approach for identifying latent creative capacities in non-specialized individuals. In contrast, CAT is designed to evaluate realized creative outputs, leveraging expert judgment to assess creativity in specialized domains such as art, design, and innovation. Its evaluative dimensions—creativity, technical skill, and aesthetic appeal—allow for subjective and context-sensitive assessments of creative performance. This methodological distinction makes CAT especially suitable for domains requiring nuanced interpretation of creativity.

TTCT and CATare integrated to enable us to capitalize on their strengths in this study. TTCT offers a robust foundation for evaluating creative potential, while CAT emphasizes the realized quality of creative outputs, aligning with our interdisciplinary focus on assessing

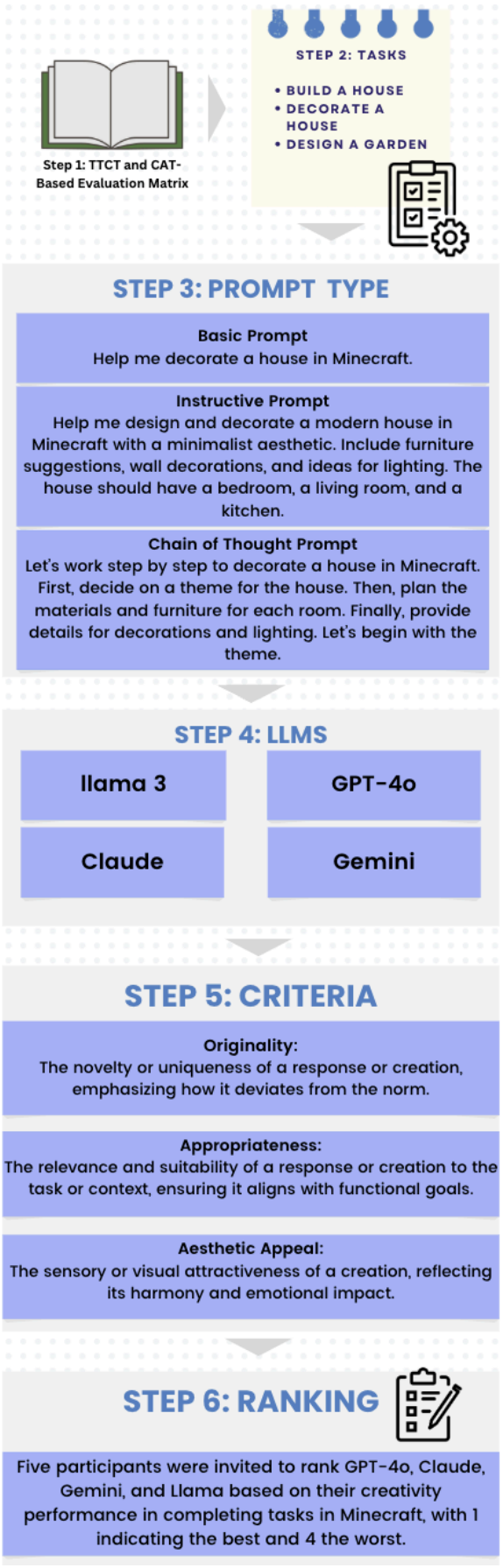


Fig. 2 Methodology Flowchart

creativity in complex and interactive contexts. The rationale for combining these frameworks is rooted in their overlapping evaluative dimensions: **originality**, **appropriateness**, and **aesthetic appeal**. Originality captures the novelty of ideas or solutions generated by AI, reflecting its capacity to produce unconventional and innovative outputs. Appropriateness examines the relevance and practicality of the AI's actions or creations within the specific task context, ensuring alignment with functional objectives. Aesthetic appeal evaluates the sensory or visual quality of the AI's outputs, including their harmony, coherence, and emotional resonance. These three dimensions collectively provide a comprehensive framework for assessing AI creativity, enabling a multidimensional analysis of potential and realized performance.

| Metric | Combined Definition |
|---|---|
| **Originality** | The novelty or uniqueness of a response or creation, emphasizing how it deviates from the norm. |
| **Appropriateness** | The relevance and suitability of a response or creation to the task or context, ensuring it aligns with functional goals. |
| **Aesthetic Appeal** | The sensory or visual attractiveness of a creation, reflecting its harmony and emotional impact. |

Table 1: Creativity Evaluation Metrics

Building on the decision to use **originality**, **appropriateness**, and **aesthetic appeal** as the core evaluation metrics, we designed three tasks to assess AI creativity within the Minecraft environment: 1. Build a House where the AI constructs functional structures with specific architectural features (Fig. 3.1) 2. Decorate a House, where the AI embellishes an existing structure based on thematic prompts. (Fig. 3.2) 3. Design a Garden where the AI arranges garden elements with attention to layout, aesthetics, and emotional alignment. (Fig. 3.3) These tasks were chosen for their relevance to common Minecraft building challenges and their ability to evaluate both structural and stylistic creativity. By spanning indoor and outdoor scenarios, they provide a balanced framework for analyzing AI performance across diverse creative contexts, capturing the essential dimensions of creativity within practical and meaningful tasks.

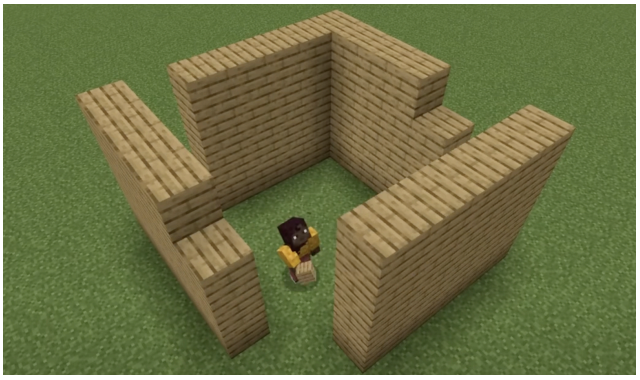

Fig. 3.1 Andy is building a house


Fig. 3.2 Andy is decorating a house


Fig. 3.2 Andy is designing a garden

Three prompt styles were employed to further explore how different instruction types influence AI performance. The Basic Prompt offered simple, open-ended instructions, such as "Help me decorate a house in Minecraft." The Instructive Prompt provided more detailed guidance, specifying requirements for the output, such as "Help me design and decorate a modern house with a minimalist aesthetic, including furniture suggestions and wall decorations." The Chain of Thought Prompt guided the AI through incremental reasoning, using step-by-step instructions like "Let us decide on a theme for the house, then plan materials and furniture for each room. Finally, provide details for lighting and decorations." This prompt variety ensured that the study captured a broad spectrum of the AI's problem-solving approaches under varying levels of complexity and specificity.

Building on the prompt styles outlined above, we selected six popular and representative architectural and interior design styles for crafting prompts: Modern Style, Scandinavian Style, Industrial Style, Rustic Farmhouse Style, Classic European Style, and Bohemian Style. These styles were chosen for their prominence in contemporary design and distinctive aesthetic and emotional characteristics, making them ideal benchmarks for evaluating AI's creative adaptability. For each style, the related keywords we used in our prompt were embedded into a Word2Vec model, a natural language processing method that maps words into a vector space. A 2D visualization of the embeddings was generated (Fig. 4, 5, 5), revealing that keywords associated with each style were clustered into distinct regions. This clustering demonstrated that each architectural style possesses a unique linguistic footprint, which can be effectively captured and utilized in our prompt. For each style, we carefully developed prompts based on the principles of the respective design, ensuring alignment with its defining features. For example, prompts for Modern Style emphasized clean lines, neutral colors, and minimalist aesthetics, while Rustic Farmhouse focused on warm textures, natural materials, and vintage elements. By testing the AI agent's performance against these prompts, we aimed to evaluate its ability to adapt its decorative choices to specific thematic and emotional styles. This approach enabled us to assess the agent's **appropriateness**, ensuring that its outputs aligned with the intended design goals and examining how creatively it interpreted and executed each style.

```
categories = [
    "Modern Style",
    "Scandinavian Style",
    "Industrial Style",
    "Rustic/Farmhouse Style",
    "Classic European Style",
    "Bohemian Style",
]

my_words = [
    # Modern Style
    "simple", "bright", "stylish", "functional", "clean",

    # Scandinavian Style
    "fresh", "tranquil", "natural", "practical", "soft",

    # Industrial Style
    "unique", "vintage", "bold", "industrial", "sturdy",

    # Rustic/Farmhouse Style
    "warm", "rustic", "relaxed", "nostalgic", "vintage",

    # Classic European Style
    "luxurious", "graceful", "noble", "refined", "ornate",

    # Bohemian Style
    "vibrant", "free", "adventurous", "eclectic", "casual",
]
```

Fig. 4 Sentimental Analysis Code: This table displays adjectives associated with different styles, which are incorporated into prompts to show how each style includes specific descriptive words in our code.

| | Metrics | Originality | | | | Appropriateness | | | | Aesthetic Appeal | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| human evaluator | task | GPT-4o | Claude | Gemini | Llama | GPT-4o | Claude | Gemini | Llama | GPT-4o | Claude | Gemini | Llama |
| 1 | build house | 4 | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 2 | 1 | 3 | 4 |
| | decorate house | 3 | 4 | 2 | 1 | 2 | 1 | 3 | 4 | 1 | 2 | 4 | 3 |
| | design garden | 2 | 1 | 3 | 4 | 1 | 2 | 3 | 4 | 2 | 1 | 4 | 3 |
| 2 | build house | 4 | 2 | 1 | 3 | 1 | 2 | 3 | 4 | 2 | 1 | 3 | 4 |
| | decorate house | 3 | 3 | 1 | 4 | 2 | 1 | 3 | 4 | 1 | 1 | 4 | 3 |
| | design garden | 1 | 2 | 2 | 3 | 2 | 1 | 4 | 3 | 1 | 2 | 4 | 3 |
| 3 | build house | 4 | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 2 | 1 | 3 | 4 |
| | decorate house | 4 | 2 | 3 | 1 | 1 | 3 | 2 | 4 | 1 | 2 | 4 | 3 |
| | design garden | 2 | 1 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 4 | build house | 4 | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 2 | 1 | 3 | 4 |
| | decorate house | 3 | 2 | 1 | 4 | 2 | 1 | 4 | 3 | 2 | 1 | 4 | 3 |
| | design garden | 1 | 2 | 4 | 3 | 1 | 2 | 3 | 4 | 1 | 2 | 4 | 3 |
| 5 | build house | 4 | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 2 | 1 | 3 | 4 |
| | decorate house | 2 | 1 | 3 | 4 | 3 | 1 | 2 | 4 | 2 | 1 | 4 | 3 |
| | design garden | 1 | 2 | 3 | 4 | 2 | 1 | 3 | 4 | 2 | 1 | 4 | 3 |
| summary statistics | | 4 | 1 | 2 | 3 | 1 | 2 | 3 | 4 | 2 | 1 | 4 | 3 |

Table 2: Table 2: Participant Rankings for AI Models Across Various Tasks. The numbers (1, 2, 3, 4) represent the first, second, third, fourth, and fifth human participants, respectively. The tasks listed correspond to different evaluation metrics, including **originality**(yellow), **appropriateness** (green), and **aesthetic appeal** (blue)

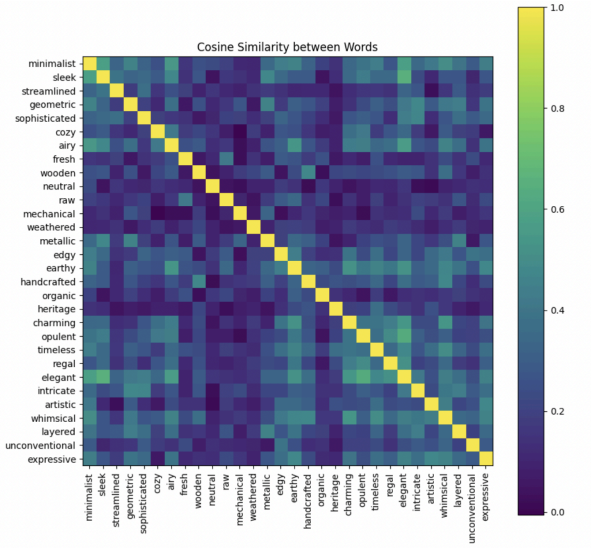


Fig. 5 Cosine Similarity Visualization: This visualization shows the strength of similarity between terms, where lighter colors (yellow) indicate higher similarity and darker colors (blue) indicate lower similarity.

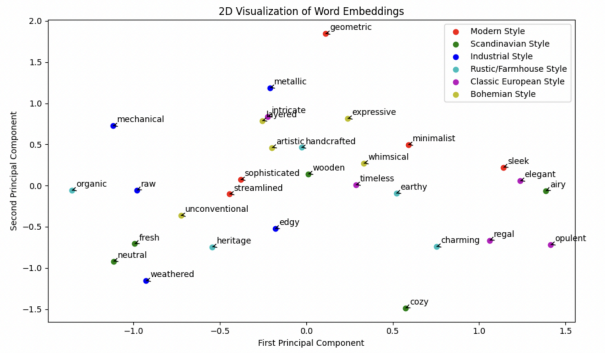


Fig. 6 Word2Vec 2D Visualization: This chart maps each word to a specific coordinate in a vector space, visualizing the spatial relationships and positions of words.

After each AI agent completed its assigned tasks, the outputs were saved for the final evaluation stage: ranking by human participants. For this, we recruited five participants, all of whom are current students at the University of Virginia. These participants were tasked with ranking the outputs based on the three evaluation metrics—Originality, Appropriateness, and Aesthetic Appeal. Participants were first provided with clear definitions of these metrics in the psychology test to ensure consistency and understanding, supplemented with concrete examples to illustrate their meanings.

Once participants fully understood the evaluation criteria, they were shown the outputs for all tasks completed by the four AI models (GPT-4o, Claude, Gemini, and Llama). For example, Participant 1 reviewed the results of each model's performance in the Build a House, Decorate a House, and Design Garden tasks, resulting in 12 outputs in total (3 tasks × 4 models). For each task, participants ranked the four models' outputs from first place (best) to fourth place (least effective) based on the specified metric. This process was repeated separately for each of the three metrics, as the definitions and criteria for Originality, Appropriateness, and Aesthetic Appeal vary and require independent evaluations.

Each participant conducted a total of 36 evaluations (12 outputs × 3 metrics), using their judgment and cross-comparisons of the outputs. The rankings aimed to capture the participants' perceptions of the models' performance across the creativity dimensions. Finally, the results were aggregated to identify collective trends in human evaluation of AI creativity. This aggregation allowed us to determine whether the participants' evaluations displayed a consistent preference pattern or were more random and distributed. The final rankings provided insights into how humans perceive the creative performance of AI agents across diverse tasks.

## Experiments and Results

The evaluation phase involved five participants, each tasked with ranking the outputs of four AI

models—GPT-4o, Claude, Gemini, and Llama—based on three creativity metrics: Originality, Appropriateness, and Aesthetic Appeal. Each participant completed a total of nine rankings, with three rankings assigned to each metric. Participants ranked the models' outputs for every metric across three tasks: Build a House, Decorate a House, and Design a Garden, assigning scores from first (best) to fourth (least effective) for each task. This process allowed participants to evaluate the performance of the AI models within each creativity dimension, providing a detailed and systematic assessment for each task. After completing all nine rankings, the results from each participant were recorded, yielding a total of 45 rankings per metric across all participants.

**Appropriateness**, defined as the relevance and suitability of a response to the task context, demonstrated the most evident trend among participants. GPT-4o consistently ranked highest in this metric, reflecting its ability to produce functional and contextually aligned outputs that closely adhered to the task prompts. Claude followed as the second-best performer, with generally well-aligned outputs, but occasionally exhibited minor inconsistencies. In contrast, Gemini and Llama were typically ranked third and fourth, respectively, as their outputs were often perceived as random or impractical. For example, in building tasks, Gemini and Llama frequently generated designs that needed more logical coherence and usability expected by human interpretations of houses and gardens.

**Aesthetic Appeal**, which evaluates outputs' visual and emotional quality, also showed a strong and consistent trend. Participants overwhelmingly favored Claude's outputs in this category, citing their visual coherence and pleasing design. GPT-4o ranked second, producing symmetrical and consistent designs that were functional but lacked the creative flair of Claude's outputs. Gemini and Llama typically ranked lower, as their designs were either overly simplistic or visually chaotic, failing to meet participants' expectations for aesthetically engaging outputs.

**Originality**, measuring the novelty and creativity of the outputs, exhibited more significant variability in participant rankings than the other metrics. Claude frequently ranked highest for its ability to balance human-like logic with creative randomness, producing outputs that participants described as innovative and comprehensible. While GPT-4o often ranked second, participants noted that its outputs were conventional and lacked distinctive originality. Gemini and Llama occasionally ranked higher due to their abstract and unconventional outputs, which some participants interpreted as creative. However, this randomness often detracted from the outputs' overall functionality and appropriateness, leading to mixed evaluations.

The rankings were consolidated by calculating the rank frequency of LLMs for each metric across all tasks, yielding a final ranking summary. For Originality, Claude ranked first, followed by Gemini, Llama, and GPT-4o. For Appropriateness, GPT-4o ranked first, followed by Claude, Gemini, and Llama. For Aesthetic Appeal, Claude ranked first, with GPT-4o, Llama, and Gemini following. These results reflect consistent trends, with GPT-4o and Claude performing well in metrics requiring practicality and aesthetic coherence, while Gemini and Llama displayed weaknesses in alignment and design sophistication.

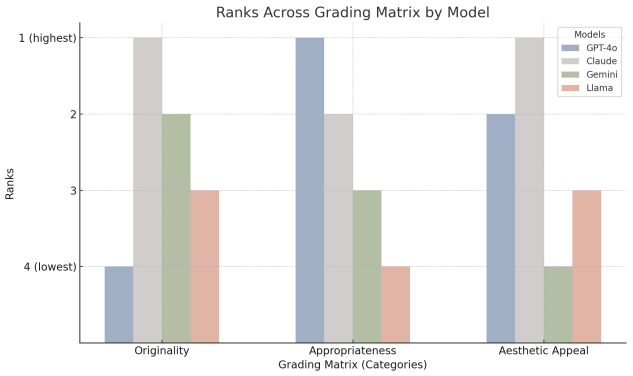


Fig. 7 Ranks Across Grading Matrix by Model：This legend summarizes the ranking of four AI agents—GPT-4, Claude, Gemini, and Llama—based on their performance across three dimensions: **Originality**, **Appropriateness**, and **Aesthetic**. The rankings reflect the aggregated participant evaluations, showing which agent ranked first, second, third, and fourth in each category.

The analysis of ranking distributions revealed clear preferences among human participants, which were further visualized through summary bar charts. (Fig. 7) These visualizations confirmed that GPT-4o and Claude consistently outperformed Gemini and Llama across most metrics. Participants favored GPT-4o for its practical and functional outputs in Appropriateness and Claude for its visually appealing and creatively balanced designs in Aesthetic Appeal and Originality. While rankings for Originality exhibited more variability, Claude's ability to combine novelty with comprehensibility gave it an edge. The findings indicate that human evaluators favored models that effectively balanced innovation, functionality, and visual quality, revealing a clear trend in participant preferences. These results highlight the strengths and limitations of current AI models in generating creative outputs for Minecraft tasks, providing valuable insights into AI creativity.

Building on the previous findings from participant feedback, where AI models demonstrated strong alignment between thematic prompts and design outputs—particularly in their ability to reflect linguistic and stylistic coherence—we conducted a focused Sentimental Analysis experiment. This experiment further evaluated the AI's ability to match the architectural styles previously used in the prompts. Conducted in Minecraft Java Edition 1.21, featuring 830 placeable blocks, the experiment showed that AI agents consistently selected blocks that were thematically appropriate and aligned with

the stylistic requirements of user-defined prompts. This capability highlights the AI's strong adaptability to thematic constraints and effective use of the block dataset to reflect aesthetic and emotional goals.


Fig. 8.1 AI's Selection on the Building Blocks

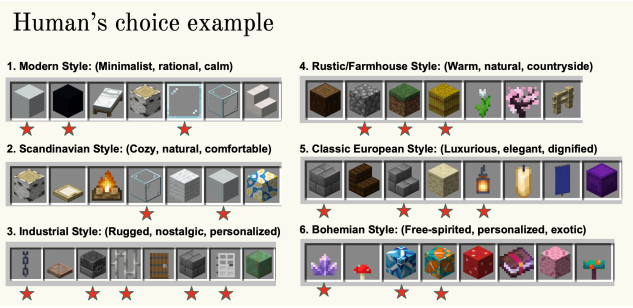


Fig. 8.2 Human's Selection on the Building Blocks

To validate the AI's performance, we conducted an additional experiment involving five human participants alongside **Claude**, one of the highest-performing agents, to test its material selection for six architectural styles: **Modern**, **Scandinavian**, **Industrial,** **Rustic**, **Classic European**, and **Bohemian**. Claude's block choices were recorded and compared to the human participants, who performed the same task. Analysis revealed a high degree of alignment between AI and human choices, with many identical block selections, as highlighted in Fig. 8.1 and Fig. 8.2, where red stars indicate blocks chosen by both the AI and humans. Given the randomness of block selection probabilities (1/830), this repeated overlap demonstrates the AI's strong ability to associate blocks with stylistic descriptors, reflecting a form of common sense in its aesthetic decision-making.

However, subtle differences emerged. AI selections tended to be more "conservative," favoring basic, universally recognized blocks for each style. In contrast, human participants showed more significant variance, often combining materials with varied textures and colors to produce more nuanced results. This distinction suggests that while AI aligns well with predefined stylistic themes, it lacks the creative flexibility and originality in human outputs in this experiment setting.

Follow-up interviews with participants further emphasized this gap. While participants were impressed by the AI's accuracy in Sentimental Matching, they noted its limitations in open-world creativity. Three experienced Minecraft players observed that the AI agents lacked the virtual contextual adaptability and imaginative complexity required to create outputs that rival human creativity. While humans could introduce innovative combinations and complex structures to fit the environment, the AI's outputs were functional but lacked depth, often reflecting safe, template-like designs. These findings highlight the AI's strengths in accurately interpreting stylistic prompts and selecting appropriate materials and its limitations in broader creative contexts and environment fitness.

## Discussion

This study highlights several key findings regarding the performance and potential of AI agents in dynamic virtual environments. First, a clear trend emerged in participant evaluations, demonstrating that different AI models excel at specific tasks. For example, GPT-4o showed strengths in appropriateness and functional alignment, while Claude stood out in aesthetic appeal and originality. However, models like Gemini and Llama lagged, often generating outputs perceived as less coherent or practical. These findings suggest good opportunities for optimization, particularly in elevating underperforming models to match the levels of more successful ones. The creative capabilities of all AI models could be further enhanced through improved task design, prompt engineering, and model fine-tuning.

Second, the study revealed the impressive ability of AI agents to perform sentimental matching in tasks involving material selection and stylistic alignment. Participants were particularly surprised by the AI's ability to interpret descriptive prompts and select blocks that visually and thematically matched architectural styles. This capability was effectively visualized through Minecraft Dynamic Virtual Worlds, demonstrating the AI's logical consistency and alignment with human expectations. Many participants expressed enthusiasm about the potential for integrating such agents into their gaming experiences, noting that AI companions could enhance motivation and creativity during gameplay. However, they also pointed out practical limitations, such as the technical skills required for setup and operation to replicate this study environment, which might hinder accessibility for non-technical users. Despite this, participants were optimistic about the AI's potential to foster creativity and engagement in gaming and other interactive tasks.

Looking ahead, the societal implications of AI agents in open-world environments are significant. These systems could serve as valuable tools for entertainment, education, and creative exploration. By providing interactive companionship and enabling collaborative creativity, AI agents could reduce feelings of isolation in solo gaming experiences while enhancing enjoyment and innovation.

Moreover, future iterations of these agents could be adapted for broader applications in open-world games, design tools, and other creative platforms, empowering users to push the boundaries of their imagination and productivity.

## Conclusion

This research provides a comprehensive benchmark of AI agents' creativity within Minecraft, demonstrating their strengths in appropriateness, aesthetic appeal, and sentimental matching while highlighting areas for improvement in originality and accessibility. By leveraging established psychological frameworks and participant feedback, the study underscores the potential of AI to enhance creativity, engagement, and collaboration in virtual environments. With further optimization, these agents hold promise as interactive tools for gaming, education, and creative industries, offering innovative ways to enrich user experiences and foster human-AI collaboration.

## Reference:

[1] Chen, Y. *et al.* (2024) *Assessing and understanding creativity in large language models*, *ar5iv*. Available at: https://ar5iv.labs.arxiv.org/html/2401.12491 (Accessed: 12 December 2024).

[2] Cseh, G. and Jeffries, K. (2019) *Apa PsycNet*, *American Psychological Association*. Available at: https://psycnet.apa.org/record/2019-20312-005 (Accessed: 12 December 2024).

[3] Fan, L. *et al.* (2022) *Minedojo: Building open-ended embodied agents with internet-scale knowledge*, *arXiv.org*. Available at: https://arxiv.org/abs/2206.08853 (Accessed: 12 December 2024).

[4] Nottingham, K. (2024) *Kolbytn/mindcraft*, *GitHub*. Available at: https://github.com/kolbytn/mindcraft (Accessed: 12 December 2024).

[5] Nottingham, K. *et al.* (2023) *DO embodied agents dream of pixelated sheep: Embodied decision making using language guided World modelling*, *arXiv.org*. Available at: https://arxiv.org/abs/2301.12050 (Accessed: 12 December 2024).

[6] PrismarineJS (2024) *PrismarineJS/mineflayer: Create minecraft bots with a powerful, stable, and high level javascript API.*, *GitHub*. Available at: https://github.com/PrismarineJS/mineflayer (Accessed: 12 December 2024).

[7] Torrance, E.P. (1966) *Apa PsycNet*, *American Psychological Association*. Available at: https://psycnet.apa.org/record/9999-05532-000?doi=1 (Accessed: 12 December 2024).

[8] Wang, G. *et al.* (2023) *Voyager: An open-ended embodied agent with large language models*, *arXiv.org*. Available at: https://arxiv.org/abs/2305.16291 (Accessed: 12 December 2024).