Creativity Research Journal

ISSN: 1040-0419 (Print) 1532-6934 (Online) Journal homepage: www.tandfonline.com/journals/hcrj20

A Rubric-Guided Multimodal Approach Using High-Capacity LLMs Provides Psychometrically Sound Creativity Assessment in Learning Games

Seyedahmad Rahimi, Hongming Li, Salah Esmaeiligoujar, Deniz Ercan & Anthony Botelho

To cite this article: Seyedahmad Rahimi, Hongming Li, Salah Esmaeiligoujar, Deniz Ercan & Anthony Botelho (04 Mar 2026): A Rubric-Guided Multimodal Approach Using High-Capacity LLMs Provides Psychometrically Sound Creativity Assessment in Learning Games, Creativity Research Journal, DOI: 10.1080/10400419.2026.2638384

To link to this article: https://doi.org/10.1080/10400419.2026.2638384

CREATIVITY RESEARCH JOURNAL
https://doi.org/10.1080/10400419.2026.2638384

Routledge
Taylor & Francis Group

A Rubric-Guided Multimodal Approach Using High-Capacity LLMs Provides Psychometrically Sound Creativity Assessment in Learning Games

Seyedahmad Rahimi, Hongming Li, Salah Esmaeiligoujar, Deniz Ercan, and Anthony Botelho

University of Florida

ABSTRACT

Creativity assessment at scale is difficult because expert ratings are resource-intensive and hard to use in dynamic settings. Large language models (LLMs) offer potential for automated assessment, yet their validity for evaluating multimodal creative artifacts in authentic educational contexts remains unexplored. Here we evaluated whether LLMs can assess human creativity in *Physics Playground*, an educational physics game where students design playable levels. We compared rubric-guided and rubric-free prompting approaches across 421 student-created levels, tested three multimodal input configurations, and examined reliability and model capacity effects using GPT and Gemini model families. Rubric-guided prompting yielded strong agreement with human expert ratings compared to rubric-free approaches (*r*s ranged from .61 to .81). Multimodal inputs combining images with structured data significantly enhanced validity compared to text-only methods. These effects were consistent across GPT-4o and Gemini 2.5 Flash. Also, single model calls achieved comparable reliability to averaged responses. Model capacity substantially influenced performance, with larger, high-capacity models (e.g. GPT-4o) consistently outperforming smaller, low-capacity variants (e.g. GPT-4o-mini). Theoretically, these findings extend creativity assessment to multimodal artifacts in authentic contexts. Practically, embedding assessment in learning games enables them to foster creativity and support STEM learning and AI literacy.

Introduction

Creativity is widely recognized as a cornerstone of human progress and as a durable 21st-century competency that is essential for driving innovation and navigating complex, technology-rich societies (Craft, 2010; Glaveanu et al., 2020; OECD, 2024; UNESCO, 2023; World Economic Forum, 2025). One of the most frequently used and agreed-upon definitions of creativity is that it involves generating products (e.g., ideas, solutions, or artworks) that are both *novel* and *appropriate* (Amabile, 2018; Kaufman & Beghetto, 2009; Kaufman & Sternberg, 2010; Plucker et al., 2004). Novelty refers to originality, while appropriateness refers to being logical, functional, practical, and valuable. Scholars have also emphasized additional qualities such as surprise, elegance, and aesthetic appeal (Cropley & Cropley, 2011; Sternberg & Lubart, 1996). More recent perspectives argue that creativity is deeply embedded in social, cultural, and technological contexts that require not only originality and utility but also inclusivity and adaptability (Beghetto & Karwowski, 2023; Runco, 2023).

From a learning perspective, recent research suggests that creativity enhances learning by enabling learners to form novel connections among concepts, a process particularly critical for mastering complex STEM topics (Kwon & Lee, 2025; Luchini et al., 2025; Syamra & Suryadi, 2025). In this context, Generative AI (GenAI) opens new possibilities for creativity; however, these opportunities are accompanied by significant challenges (Beaty et al., 2022; Chaudhry & Kazim, 2022). GenAI tools can generate text, music, and visual art that often rival human outputs (Anantrasirichai et al., 2025; Heigl, 2025). Educators and creative professionals can now experiment with GenAI as a collaborative partner to spark ideas and expand design spaces (Cai & Gao, 2025; Urmeneta & Romero, 2025; Wingström et al., 2024).

Yet, questions remain about whether GenAI genuinely "creates" or merely recombines human-made artifacts. Critics argue that Large Language Models (LLMs) lack lived experiences, emotions, and contextual understanding (often considered essential to human creativity; Lockhart, 2025). Irmak (2024) introduces the concept of "authorless artifacts" and suggests that

CONTACT Seyedahmad Rahimi srahimi@ufl.edu College of Education, School of Teaching and Learning, University of Florida, 2821 Norman Hall, PO Box 117048, Gainesville, FL 32611, USA

many GenAI outputs lack intentional authorship and challenge traditional notions of creative ownership. Similarly, Runco (2023) further contends that LLMs generate artifacts via processes fundamentally different from humans, resulting in "artificial creativity" rather than genuine creativity. Additionally, Koivisto and Grassini (2023) showed that although LLM-based chatbots outperformed the average human in a divergent thinking task, the most creative human responses still surpassed the best AI outputs in terms of originality and overall creativity.

Highlighting the enduring value of human ideation in guiding LLMs outputs, Seli et al. (2025) showed that AI-generated images based on prompts crafted by professional artists were rated as more creative than those generated from prompts written by an AI chatbot or novice users. Thus, while human – AI collaboration holds promise, scholars caution that creativity should not be outsourced to AI but should remain grounded in human judgment and agency (Haase & Pokutta, 2024; Wingström et al., 2024).

In educational contexts, these opportunities and challenges are particularly salient. GenAI can act as a creative catalyst that inspires and scaffolds learners' ideas, yet over-reliance without human oversight may reduce critical thinking (Zhai, 2024) and originality (Rahimi et al., in press). Ethical issues such as bias, lack of representation, and digital inequity further threaten fairness and widening social divides (Rahimi et al., 2026; Dieterle et al., 2024). Therefore, AI literacy (i.e., understanding how AI works, its limits, and how to use it responsibly) has become a foundational skill (Rahimi et al., 2026; Holstein & Doroudi, 2022). For example, Gu and Ericson (2025) propose a tripartite model of AI literacy (i.e., functional, critical, and sociocultural) that fosters reflective and ethical engagement in creativity education. Functional literacy involves effectively using AI tools, enhancing creative fluency; critical literacy promotes questioning assumptions and ethical reflection to deepen originality; and sociocultural literacy addresses bias, representation, and equity, encouraging awareness of AI's broader societal impact. Together, these dimensions prepare learners to use AI responsibly and think creatively. Additionally, emerging frameworks of "creativity augmentation" suggest that AI should not replace human imagination but instead amplify learners' capacity to ideate, reflect, and co-create in digital learning environments (Beghetto & Karwowski, 2023). AI-literate learners who use AI within frameworks such as creativity augmentation can avoid the pitfalls of overreliance while leveraging its potential to enhance their creativity.

From a workforce perspective, these issues are pressing. As AI-based automation increasingly replaces routine tasks, the value of human creativity, complex problem-solving, and adaptability is growing across industries (UNESCO, 2023; World Economic Forum, 2023). Employers are prioritizing innovation, critical thinking, and collaboration as essential skills for future-ready professionals (OECD, 2024). Preparing students for such an AI-saturated job market requires not only fostering creativity but also developing valid ways to assess it (Rahimi & Shute, 2021a, Shute & Rahimi, 2021; Shute & Wang, 2016).

Before developing approaches to foster human creativity or advance human – AI co-creativity, it is essential to establish robust methods for using LLMs to assess creativity. Although the creativity literature offers a nuanced understanding of what creativity is, assessing this essential competency at scale remains a persistent challenge. Traditional creativity assessments (e.g., The Torrance Test of Creativity Assessment or Alternative Uses tests) are resource-intensive and suffer from low ecological validity to capture dynamic and situated nature of creative processes (Rafner et al., 2022, Rahimi, 2023; Rahimi & Shute, 2024). In contrast, emerging LLM-based methods offer opportunities for automated, unobtrusive, and objective assessments of creativity, especially in ecologically valid environments such as digital games (Rafner et al., 2022; Rahimi, 2023). LLMs have shown promise in this area already (more about this in the next section). Nonetheless, a critical challenge remains: how to systematically validate and refine LLM-driven assessments in echologically valid environments to ensure validity, reliability, fairness, and alignment with human values (Holmes & Porayska-Pomsta, 2023; Rahimi & Shute, 2024).

Digital games represent a particularly promising medium for both assessing and fostering creativity in highly ecologically valid environments (Rahimi & Shute, 2021a, 2024, 2021b; Qian & Clark, 2016). Games afford at least two complementary modes of creative expression: gameplay, where players engage in creative problem-solving through interaction with game elements, and *game design*, where learners construct new levels or artifacts that can serve as direct expressions of creativity. Work on AI in education highlights the potential of AI-driven systems to personalize learning experiences and adaptively support learners (Holstein & Doroudi, 2022). Other research highlights how GenAI can support creative and learning processes (Zhang & Xu, 2025) while also transforming how creativity is assessed and studied in the age of AI (Acar, 2025).

In this study, we address one issue at the intersection of creativity, GenAI, and digital learning games: *the use of LLMs to assess the creativity of human-generated game*

levels in a sandbox game. To address the objective of this study (i.e., assessment of game levels using LLMs), we faced a challenge. That is, due to the generative nature of LLMs, the response of these models may be variable each time one requests something from the models via a prompt (Achiam et al., 2023). In other words, some variation in LLMs' responses is natural. However, we can do something to reduce this variability compared to the ground truth (i.e., humans' ratings) in LLMs' responses or by changing the hyperparameters (e.g., models' temperature). Therefore, we hypothesize that to get better results (e.g., creativity assessment estimates that are closer to human ratings' estimations), we should follow the same method we usually follow for rating qualitative data by humans – using rubrics (Brookhart, 2013). When humans rate qualitative data, a rubric can help bring their ratings closer to each other and lead to higher inter-rater reliability. Thinking of an LLM as a third rater, using rubrics to get a similar rating to humans from an LLM is warranted. In this study, we examine the effects of using rubrics on creativity assessment validity across GPT-4o family models and Gemini 2.5 Flash.

Findings from this study have implications for advancing future research on creativity assessment using LLMs in ecologically valid environments such as digital games. Moreover, this study provides insights for researchers and practitioners across disciplines interested in designing learning environments that not only assess but also enhance creativity while supporting students' STEM learning and engagement. Next, we review the relevant literature concerning this study.

Background

Automated creativity assessment

In the field of creativity research, a wide range of automated approaches have been developed to evaluate creative artifacts, including text-based responses, visual products, and interactive digital creations. One foundational category of methods is stealth assessment, which unobtrusively embeds assessment into digital learning environments and collects evidence of competencies (e.g., creativity) without disrupting the learner's experience (Rahimi & Shute, 2024; Shute, 2011). Grounded in the evidence-centered design framework (ECD; Almond et al., 2015), stealth assessments draw inferences about learner competencies from behavioral interactions and generate immediate, tailored feedback. For instance, Shute and Rahimi (2021) designed and validated a stealth assessment of creativity in the sandbox game *Physics Playground* (Shute et al., 2019), which inferred creativity from gameplay indicators such as fluency, flexibility, and originality. Stealth assessment has also expanded beyond game-based environments. Rahimi an colleagues (2024) developed and validated a stealth assessment of creativity within *EarSketch*, a music-remixing and programming environment. Their findings showed a moderate, significant correlation between stealth creativity scores and expert evaluations ($r = .47$), demonstrating acceptable convergent validity and highlighting the feasibility of unobtrusive, ecologically valid creativity assessment.

Advances in Natural Language Processing (NLP) have introduced additional automated methods for evaluating creative products, particularly text-based responses. Early transformer-based encoder models such as RoBERTa and XLM-RoBERTa have shown the capacity to capture nuanced aspects of creativity in written artifacts. Goecke et al. (2024), for example, used XLM-RoBERTa to assess scientific creative thinking responses in German and found strong correspondence with human originality ratings ($r = .80$). Similarly, DiStefano et al. (2025) demonstrated that fine-tuned NLP models (including RoBERTa and GPT-2) can reliably score metaphor creativity, outperforming traditional semantic-distance approaches and generalizing to unseen prompts. These studies illustrate the value of transformer-based NLP models for automated, domain-specific creativity assessment, even though they primarily focus on constrained textual outputs.

Building on these foundations, recent advances in LLMs have opened new possibilities for more flexible and scalable creativity assessment. LLMs can assess open-ended, divergent responses with a level of nuance previously difficult to achieve. Zhao et al. (2024) used LLMs to automate creativity scoring for student-generated text. Acar et al. (2024) introduced MOTES, an LLM-powered system for assessing originality in elementary students' creative responses elicited through game-like prompts (e.g., "What would be a surprising use for a ball?" or "When I got on the school bus, I saw . . . "). MOTES demonstrated strong correlations with human ratings across multiple task types (rs = .79–.91). Organisciak et al. (2023) similarly reported that GPT-3 and GPT-4 could reliably automate divergent thinking scoring, achieving high alignment with human evaluations ($r = .81$). In addition to scoring creativity, LLMs can model dynamic cognitive processes: Hadas and Hershkovitz (2025) demonstrated that GenAI systems can score Alternative Uses Test (AUT) responses while modeling serial order effects and capturing longitudinal changes in fluency and flexibility. These studies show that LLMs can scale creativity assessment across diverse text-based tasks while maintaining strong agreement with expert raters.

Beyond accuracy, researchers have also begun to explore interpretable and explainable AI approaches to creativity assessment. Haim et al. (2024) examined the use of interpretable AI for evaluating creativity in short narrative texts. Using forma mentis networks – semantic and emotional graphs representing conceptual associations – they compared how human raters, GPT-3.5, and machine-learning models (e.g., XGBoost) predicted creativity ratings in human- and AI-generated stories. They found that although LLMs approximated human judgments, they relied on different feature-importance patterns (e.g., emphasizing emotional rather than structural cues) and evaluated their own outputs differently from human-generated texts. This work highlights the value of explainable, network-based methods for understanding how AI systems reason about creativity. However, these approaches remain limited to text. In the visual domain, Luthra (2025) introduced TraitSpaces, a psychologically grounded, interpretable framework for assessing visual creativity using GPT-4.1 and CLIP embeddings. TraitSpaces models affective, symbolic, and ethical qualities in artworks and supports trait-aware co-creation, offering a shared conceptual language through which humans and AI can communicate about creative attributes. However, despite their contributions, these approaches primarily operate on static textual or visual artifacts and have not been tested in authentic, ecologically valid learning environments where creativity unfolds dynamically.

The present study extends automated creativity assessment into a dynamic, multimodal, and highly interactive learning environment and addresses a gap in prior work that has largely focused on static text- or image-based artifacts. Specifically, this study employs rubric-guided prompting with GPT and Gemini models to evaluate the creativity of complex, student-generated game levels within an educational physics game.

Various ways to optimize LLMs behavior

LLMs, such as GPT models, are trained on vast datasets using extremely powerful computing systems that are not publicly accessible. These models usually have many parameters (e.g., GPT 3 has ~175 B parameters). These characteristics of LLMs make retraining or fine-tuning them with new data computationally expensive and resource-intensive (Bowman, 2023; Chen et al., 2021). Consequently, there are two dimensions we can maneuver on to direct LLMs' responses to what we expect: Context Optimization and LLM Optimization (see Figure 1). These two dimensions include four primary methods to adapt their behavior after training (Neves, 2024): prompt engineering, fine-tuning, retrieval-augmented generation (RAG), and a combined approach using all of these methods. Other emerging dimensions involve reasoning effort, such as enabling or constraining a model's internal thinking processes (Mondorf & Plank, 2024), and tool use or web search capabilities, which extend LLMs' ability to access and integrate external information sources in real time (Singh et al., 2025). Earlier studies show that using the three approaches (i.e., prompt engineering, fine-tuning, and RAG) in combination substantially decreases hallucination while strengthening relevance and the quality of models' output (Li et al., 2024; Niu et al., 2023; Shuster et al., 2021).

Prompt engineering, the most efficient and widely used method, involves crafting precise input prompts to guide the model's output without modifying its parameters (i.e., retraining). With this method, the LLM can

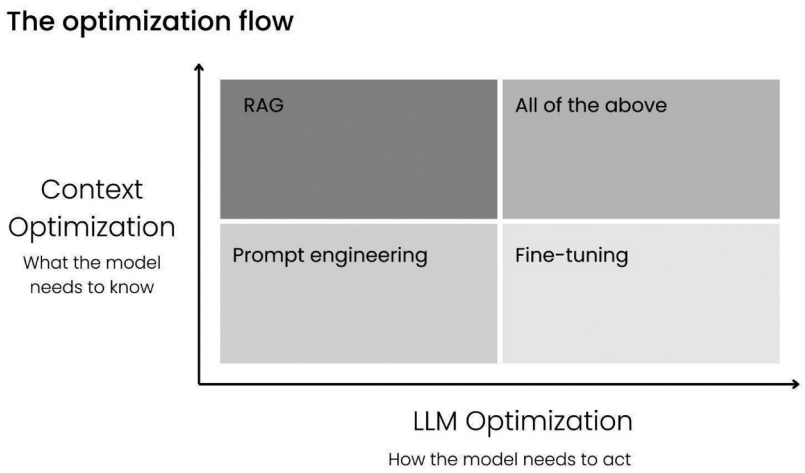


Figure 1. The Optimization flow for large language models (LLMs) (adapted from OpenAI).

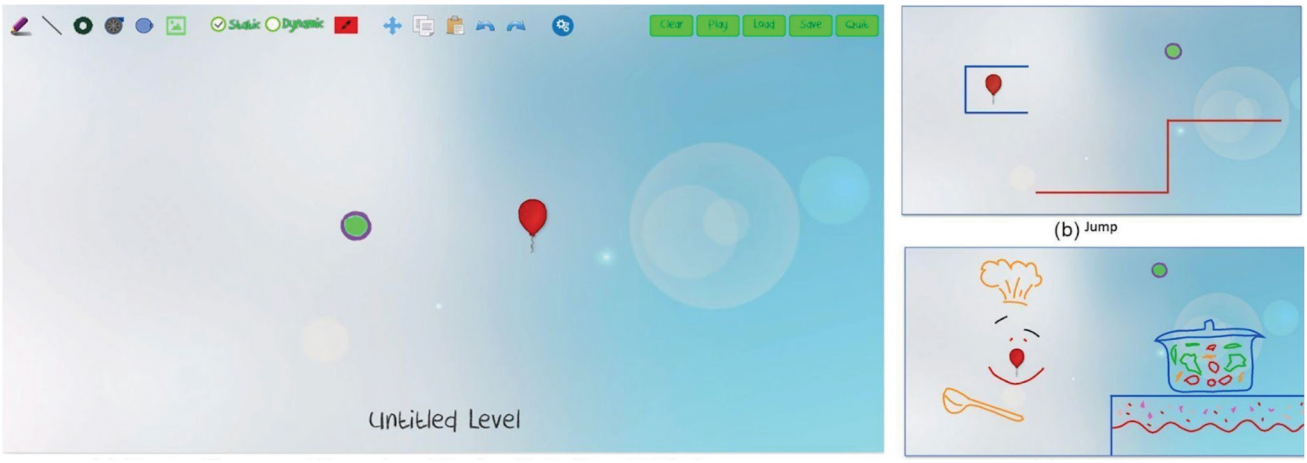
produce reasonable results using just a few examples, eliminating the need for extensive data (e.g., in our case, we don't need to show many game levels to the model). Fine-tuning, while powerful for customizing the model to specific domains, requires substantial resources, access to the model's internal parameters, and carefully curated datasets, making it costly. RAG enhances the model by dynamically integrating external knowledge sources (e.g., context-specific data) to provide updated information but requires additional infrastructure (i.e., coding and access to new data or information). Finally, a combined approach, often referred to as "All of the above," stacks these methods to maximize performance and mitigate individual shortcomings. For instance, prompt engineering can provide a foundational structure; RAG can supply context-rich updates; and fine-tuning can refine output consistency. This layered strategy is especially effective in high-stakes or complex applications, where both accuracy and reliability are essential. In this study, we used prompt engineering for its flexibility and cost-effectiveness to align the model's responses with our objectives.

Within the prompt engineering method, in this study, we systematically examined how a rubric can change the model's behavior for assessment of human-generated game levels. We also examined how calling the model several times and averaging the responses enhance the reliability of the model's response compared to only calling the model once. To further manipulate the LLM's response we examined how adding more information can affect the LLM's response (i.e., text explanation of what is included in the game level and what is the theme of the level; this information was also generated by an LLM). Finally, we also examined how different models will perform when it comes to using the same prompt we came up with.

Current study

This research focuses on the use of LLMs to assess the creativity of human- designed game levels in *Physics Playground*. *Physics Playground* is a 2-dimensional game created for 8th and 9th graders targeting Newtonian physics understanding. The goal of the game is to hit a red balloon with a green ball achieved by (a) drawing simple physics machines (i.e., ramp, springboard, lever, and pendulum) and objects (e.g., weights) in Sketching levels or (b) manipulating sliders (e.g., air resistance, gravity, mass of the ball) in Manipulation or simulation-like levels.

Many popular games have a level editor or a "create" mode where players can create their own game levels (e.g., *Little Big Planet*, *Minecraft*, *Portal 2*). This feature, as Gee (2005) indicates, follows a constructionist approach and allows players to be a part of game design and enhance their intrinsic motivation for gameplay. These games allow players to be creative through the availability of multiple tools and virtually endless possibilities. *Physics Playground* includes a level editor in which non-technical users (e.g., students and teachers) can create their own levels by drawing objects (e.g., lines, shapes, or other objects) on the screen. Starting with an empty stage (Figure 2(a)), students can place the ball and the balloon anywhere on the screen, and draw any number of obstacles between them. There are endless opportunities for students to show their creativity in this environment.



(a) Physics Playground Level Editor (c) Invisible Chef

Figure 2. Level editor (a); a game level with low creativity (b); a game level with high creativity (c).

Game levels in *Physics Playground* are stored in JSON (JavaScript Object Notation, 2023) format (Rahimi et al., 2023). JSON is a lightweight, text-based, language-independent data interchange format that is easy for humans to read and write and easy for machines to parse and generate. This structure aligns well with the way information is typically organized and conveyed in natural language (Ferrucci et al., 2010).

Given the multimodal nature of game levels, combining structured JSON data with visual representations, LLMs present a promising approach for automated creativity assessment. However, the psychometric quality of LLM-based creativity assessment in ecologically valid educational game environments remains underexplored. Key methodological questions persist regarding optimal prompting strategies, the role of different input modalities, scoring consistency across multiple model calls, and the extent to which findings generalize across model architectures. This study addresses these gaps by systematically evaluating LLM-based creativity assessment in *Physics Playground's* level editor. To that end, we address the following research questions:

RQ1) Which prompting approach produces creativity scores that best align with human expert ratings (validity)? We compare rubric-guided versus rubric-free prompts and three multimodal input combinations. To test whether findings reflect method robustness or model-specific biases, we validate across both GPT and Gemini architectures.

RQ2) How consistent are LLM-generated creativity scores (reliability)? Given the inherent stochasticity of language models, we examine whether aggregating multiple independent calls improves scoring reliability.

*RQ3) How does model capacity affect assessment performance (*efficient *scalability)? We compare models with varying computational requirements to understand cost-performance trade-offs for scalable implementation.*

The findings of this study can be used to facilitate personalized learning and creativity experiences. By understanding each learner's creative strengths and weaknesses, the system can adapt the game environment to provide tailored creativity support and foster individual growth and engagement. Next, we discuss the Methodology we used to address our research questions.

Method

In this study, we used two existing datasets of game levels from Rahimi (2020). The first dataset (DS1) consisted of game levels ($n = 30$) that were designed by the researchers and *Physics Playground* game designers. These levels were categorized into three creativity groups: Low, Medium, and High, with 10 levels in each category. The categorization was based on a structured creativity rubric, which evaluated aspects of a game level in *Physics Playground* such as elaboration (lines and meaningful objects), originality, aesthetics, humor or surprise, and title creativity (Table 1).

The second dataset (DS2) consists of 421 game levels that were created by college students during the original experimental study described in Rahimi (2020)Rahimi & Shut (2021a). Participants ($n = 114$; $M_{age} = 26.25$, $SD = 8.06$) included 54% females, 46% males; 47% undergraduate students; 53% graduate students, from various ethnicities with the majority of them as White (47%), Asian (16%), and Hispanic (15%).

Table 1. Creativity scoring rubric for levels in *Physics Playground* (adapted from Rahimi & Shute, 2021a).

Criteria	Score options	Range
Relevance	Can it be solved? (Screening Criterion) 0 = Unsolvable, do not score other variables, 1 = Solvable, continue scoring other variables	0 or 1
Line Elaboration	Is it well elaborated? (Possible scores: 0, 1, 2), Lines: If 0 lines are on the screen = 0, If 1 to 5 lines = 1, If more than 5 lines = 2	0, 1, or 2
Meaningful Objects Elaboration	Are meaningful objects well elaborated? 0 = No meaningful objects, 1 = 1–5 objects, 2 = 5+ real-life objects	0, 1, or 2
Originality	Is it original relative to existing levels? 0 = Almost identical to an existing level, 1 = Has some similarities, 2 = Very dissimilar and unique	0, 1, or 2
Aesthetics	Is it aesthetically pleasing? 0 = Poor visuals; 1–2 colors, 1 = Plain; 3–4 colors, 2 = Excellent visuals; 4+ colors	0, 1, or 2
Humor/Surprise	Does it create humor or surprise? 0 = None, 1 = Somewhat humorous/surprising, 2 = Very humorous/surprising	0, 1, or 2
Title's Creativity	Is the title creative and related to the level? 0 = No connection, 1 = Weak/descriptive, • 2 = When one or more than one of the things below were true º a remote association was used (If the title is related to the level but not in an obvious way) º if an analogy was made o or if the title helps the reader understand the level that was not possible without the title º or if the title creates a sense of surprise º or if the title makes the reader smile/laugh	0, 1, or 2

In Rahimi (2020) study, participants were first introduced to the *Physics Playground* level editor and received a tutorial explaining how to use its core features. Following the tutorial, participants were given 120 minutes to design as many game levels as they could. All participants, regardless of condition, were explicitly instructed to "be creative and try to create levels that you think no one else will create." This directive aimed to encourage divergent thinking and personal expression within the design constraints of the tool. Importantly, participants were also instructed to ensure that the levels they created could be solved by adding a functional requirement to the creative task. At the end of the session, each participant reviewed their own designs and selected the four levels they considered to be their most creative. These self-selected levels were then evaluated by two human raters using a structured creativity rubric (Table 1).

Creativity assessment rubric

The rubric we used (Table 1) to assess the creativity of game levels includes seven criteria.

See Figure 3 which illustrates six student-made game levels scored for creativity using the rubric. The seven criteria, identified from the literature, include relevance or solvability, line elaboration, meaningful objects elaboration (e.g., a flower is meaningful while a line may not be), originality, aesthetics, humor or surprise, and level title creativity. The maximum possible overall creativity score is 13. Using GPT-4o's API and via detailed prompt engineering we acquired the seven scores. In this process, we sent both the JSON file and the image of the game level to GPT-4o. Then, we conducted a correlational analysis to examine the convergent validity of the scores produced by GPT-4o.

Systematic process of examining the prompt engineering

To investigate the feasibility and consistency of using LLMs to evaluate creative level design, we conducted the analysis in three phases. In the initial phase, we randomly selected 30 diverse *Physics Playground* game levels (low, medium, and high) from the DS1. The goal of this phase was to explore whether GPT-4o could

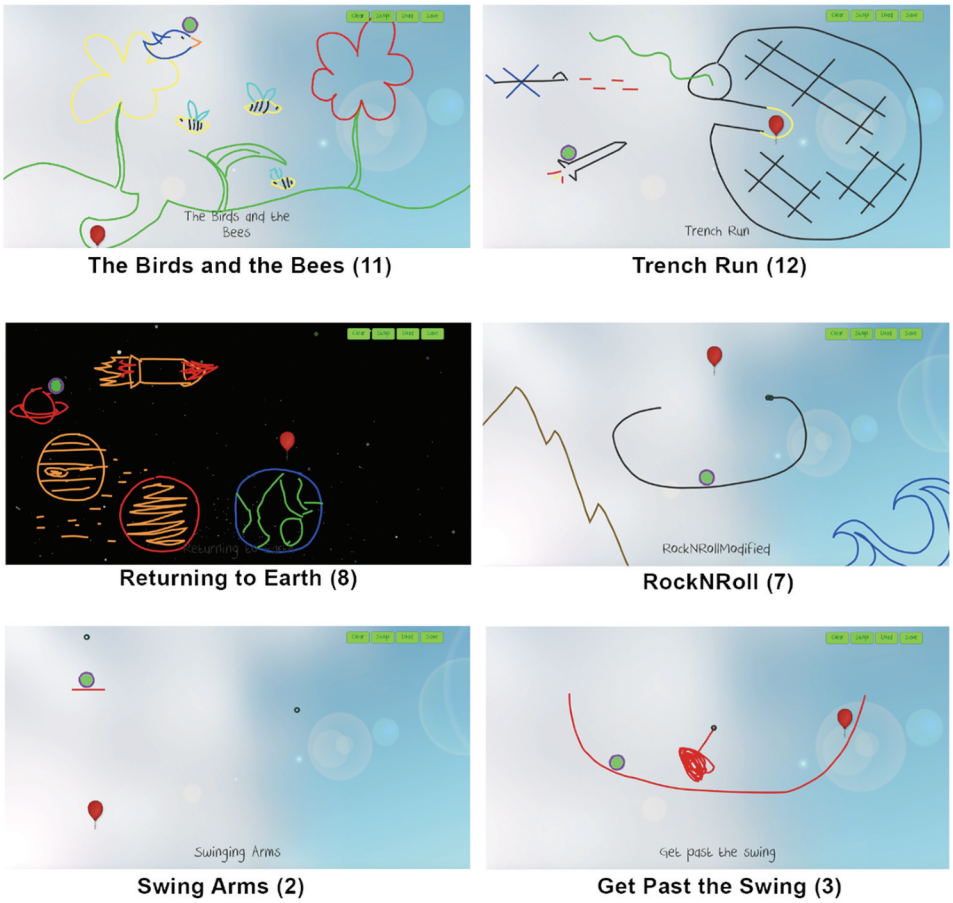


Figure 3. Six student-made game levels from DS2 with their human-rated creativity scores (max = 13).

provide ratings comparable to human evaluations using detailed prompt engineering. Each level was scored by two human raters using the creativity rubric and then rated independently by GPT-4o. We examined correlations between human and AI-generated scores to determine the convergent validity of GPT-based, automated assessment in this study's context.

Figure 4 illustrates the structure of the prompts we crafted for assessment purposes: *context* includes a detailed explanation of the game and the game mechanics, and explanation of the level editor and how game levels get saved in JSON format; detailed explanations of the *rubric* and the sub-facets with examples for each facet; target game level includes the JSON file of the game level and the image of the level which was saved within the level editor; finally, the *task* includes the request (i.e., detailed explanation of what is needed which is assessing the game levels' creativity using the given rubric), and the format we need that results in (e.g., JSON or CSV).

In the second phase, we expanded our dataset by adding 31 more levels (randomly selected from the student-made levels from DS2) to the initial 30 (n = 61). In this phase, we compared the GPT-4o's performance when using a rubric-based prompt versus a generic prompt without the rubric. Next, we randomly selected 60 new game levels from DS2 that had not been used in the prior phases. In this phase, we exclusively employed the rubric-based approach to evaluate the levels with GPT-4o. This allowed us to further validate the consistency and generalizability of rubric-guided AI assessments on a fresh dataset.

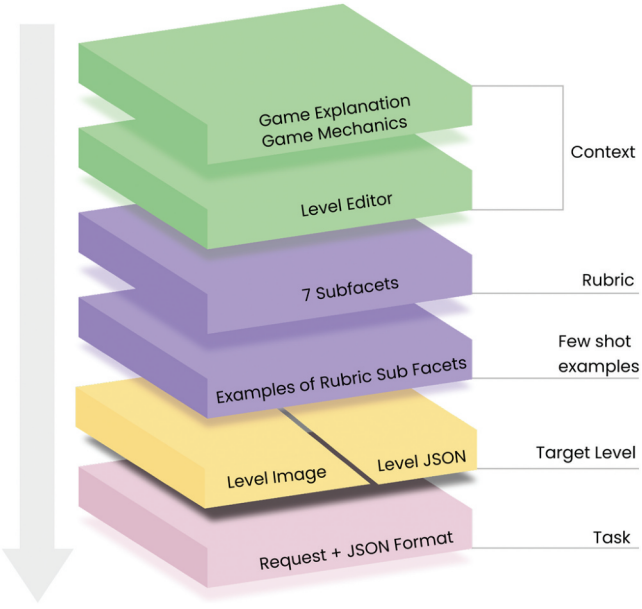


Figure 4. Prompt engineering structure for creativity assessment.

Building on the preliminary findings from the smaller datasets (DS1 and DS2), we conducted our analysis on a more comprehensive dataset (DS3, n = 421) to address our research questions. This dataset represents the complete collection of student-created game levels from the Rahimi (2020) study, of which the previously used DS2 was a subset. This larger scale allows for a more robust and generalizable evaluation of the LLM's assessment capabilities. To address RQ1, we systematically compared three prompting methods that varied in their multimodal input configurations. M1 provided JSON structure and level image simultaneously. M2 included an intermediate step in which the model first generated a textual description of the image, then used this description alongside JSON for scoring. M3 excluded the image, using only JSON and the model-generated description. These three methods allowed us to isolate the contribution of visual information (M1 vs. M3) and test whether explicit verbalization of visual content affects scoring validity (M1 vs. M2). Table 2 details the rationale for each configuration. To test whether findings reflected genuine methodological differences or model-specific artifacts, we evaluated all three methods using two architecturally distinct models: GPT-4o and Gemini 2.5 Flash. GPT-4o is based on OpenAI's proprietary multimodal architecture, while Gemini 2.5 Flash is built on Google's separate multimodal foundation, with different training data, alignment strategies, and inference mechanisms. Convergent results across both architectures would indicate method robustness independent of specific model implementation.

Model parameters and implementation Details

To ensure reproducibility and scoring consistency, we configured specific Application Programming Interface (API) parameters for model inference. These settings were chosen to maximize output determinism and structural adherence for our evaluative task, where consistency across repeated assessments is critical.

For all GPT-based experiments, we used temperature = 0.5 (default: 1.0) to reduce stochastic variation in scoring, and response_format = {"type": "json_object"} to enforce syntactically valid JSON output containing all seven creativity dimensions' scores. We tested three model variants: gpt-4o (primary), gpt-4o-mini (mid-capacity), and gpt-4.1-nano (low-capacity). All other parameters remained at default values (top_p = 1.0, frequency_penalty = 0, presence_penalty = 0), as these primarily influence generative text diversity rather than structured evaluation tasks.

Table 2. Definition and rationale of the three prompting methods compared.

Method	Description	Rationale/Hypothesis
M1	Image + JSON	• To establish a baseline for a direct multimodal assessment where the LLM must interpret both visual and structured data simultaneously.
M2	Image + JSON + Description	• To test if adding an explicit, AI-generated text description of the image enhances performance by simplifying the visual analysis task for the final scoring prompt.
M3	No Image (JSON + Description)	• To serve as a crucial control condition to isolate the specific contribution of the visual modality. This tests if the model can achieve high validity using only non-visual information.

For Gemini 2.5 Flash experiments, we set response_mime_type = "application/json" and provided a response_schema defining the required seven-dimension structure, functionally equivalent to GPT's JSON formatting approach. Additionally, we configured safety_settings to BLOCK_NONE across all categories to prevent the model's content filters from incorrectly flagging abstract game elements (e.g., level titles, geometric shapes) as harmful content, which would block valid scoring outputs. Temperature and sampling parameters remained at default values (temperature = 1.0, top_p = 1.0), as the strict schema enforcement provided sufficient output consistency.

Next, we present our findings through the iterative process explained above.

Results

To address the first research question related to convergent validity, we first used DS1 consisting of researcher-created game levels ($n = 31$) as our "training" dataset. These levels included game levels with low, medium, and high levels of creativity. The average overall creativity score from human raters (using the rubric in Table 1) for DS1 levels was 8.32 ($SD = 3.31$). The average overall creativity score from GPT-4o for DS1 levels was 7.94 ($SD = 3.78$) when using the rubric and 8.16 ($SD = 2.02$) without the rubric. Additionally, we conducted correlation analyses to examine validity of GPT-4o scores compared to human ratings under both *with* and *without* rubric conditions (to address RQ 1) in DS1 (see Table 3). Results showed a moderate correlation under the without rubric condition ($r = .64$) and a strong correlation under the with rubric condition ($r = .84$).

Building on the promising results from DS1, a second existing dataset (DS2, $n = 30$) was used with students-created game levels. The nature of DS2 allowed us to test the assessment on the type of game levels that we expect to see in the future from novice *Physics Playground* users (e.g., students). The average overall creativity score from human raters for DS2 levels was 7.65 ($SD = 3.01$). We conducted an independent sample t-test to ensure two DS1 and DS2 were statistically similar. The t-test results showed no significant difference between the datasets (t (59) = 0.83, $p = .41$, $d = .21$). This suggests that DS1 and DS2 are comparable. The average creativity score from GPT-4o for DS2 levels was 8.13 ($SD = 1.97$) when using the rubric, compared to 8.17 ($SD = 1.97$) when rubric was not used. The correlational analyses examining the validity of GPT-4o scores compared to human ratings under both *with* and *without* rubric conditions in DS2 were in agreement with the results from DS1 correlational analyses. That is, we found a moderate correlation under the without rubric condition ($r = .56$) and a strong correlation under the with rubric condition ($r = .78$). When we combined DS1 and DS2, we found similar results with a moderate correlation under the without rubric condition ($r = .61$) and a strong correlation under the with rubric condition ($r = .81$). Lastly, we reported the Root Mean Square Error (RMSE) for total creativity scores, comparing human ratings (ground truth) with GPT-4o scores. The with-rubric condition ($RMSE = 1.85$) consistently showed lower error than the without-rubric condition ($RMSE = 2.53$) across both datasets. Table 3 highlights stronger agreement between GPT-4o and human scores under rubric conditions, except for Title's Creativity (TC) (rs ranging from −.19 to .26).

Table 3. Correlation of rubric criteria with and without rubric across data sets.

	n	S (%)	LE	OE	O	A	H&S	TC	TS	Total RMSE
DS1 with rubric	31	.87	.53	.84	.68	.67	.80	.05	.84	1.81
DS1 without rubric		.90	.14	.73	.40	.30	.71	−.19	.64	2.46
DS2 with rubric	30	.90	.54	.84	.65	.70	.57	.24	.78	2.03
DS2 without rubric		1	.44	.48	.53	.35	.32	.26	.56	2.47
DS1 + DS2 with rubric	61	**.88**	**.47**	**.82**	**.65**	**.69**	**.72**	**.05**	**.81**	**1.85**
DS1 + DS2 without rubric		.96	.18	.61	.46	.31	.57	−.09	.61	2.53

Note: DS = Dataset; LE = Line Elaboration, OE = Object Elaboration, H & S = Humor & Surprise, O = Originality, A = Aesthetics, TC = Title's Creativity, TS = Total Score, S = Solvability. Bold = Superior Performance of GPT-4o *with* rubric.

We conducted a systematic comparison using the full dataset (DS3, $n = 421$) to expand addressing RQ1. We compared three multimodal input configurations (M1: JSON + Image; M2: JSON + Image + Description; M3: JSON + Description) across two LLM families: GPT-4o and Gemini 2.5 Flash.

For GPT-4o (see Table 4), M1 achieved the highest correlation with human ratings ($r = .74$, 95% CI [.69, .79]), followed by M2 ($r = .71$, CI [.65, .77]) and M3 ($r = .68$, CI [.62, .74]). The difference between M1 and M3 was statistically significant ($\Delta r = .06$, $p < .05$), while M1 and M2 did not differ significantly ($\Delta r = .03$, $p = .18$). Dimension-level analyses showed consistent patterns, with Object Elaboration showing the strongest correlations ($r = .73$ for M1) and Title Creativity the weakest ($r = .35$ for M1).

For Gemini 2.5 Flash (Table 5), correlations were systematically lower across all methods: M1 ($r = .57$, CI [.51, .63]), M2 ($r = .54$, CI [.47, .61]), and M3 ($r = .48$, CI [.41, .55]). Direct comparison revealed that GPT-4o outperformed Gemini 2.5 Flash by .17 to .20 correlation points across all three configurations which represents a 23–29% relative improvement in explained variance. Both models exhibited similar dimension-level patterns, with Object Elaboration and Line Elaboration yielding higher correlations than Title Creativity and Humor & Surprise. Notably, GPT-4o maintained correlations above $r = .68$ across all methods, while Gemini's highest correlation (M1, $r = .57$) fell below GPT-4o's lowest (M3, $r = .68$). Given GPT-4o's consistently higher validity across all input configurations and creativity dimensions, we selected it as the base model for subsequent reliability and model capacity analyses (RQ2 and RQ3).

To address RQ2 and RQ3, which examine scoring reliability and model capacity effects respectively, a methodological decision was required. While M1 demonstrated the highest validity in RQ1, we selected M2 for reliability testing because its two-step process – first generating a text description of the game level, then using that description to generate scores – introduces two distinct sources of stochastic variation. This architectural feature makes M2 more susceptible to inconsistent outputs than M1's single, direct judgment. Consequently, M2 serves as a more sensitive test case: if aggregation can stabilize M2's noisier process, similar or greater benefits for M1 can be inferred.

For RQ2, we compared single model calls versus averaging three independent calls. Initial exploration on a subsample of 60 levels showed identical correlations ($r = .75$), prompting analysis on the full dataset ($n = 421$).

The results of this scaled analysis are presented in Table 6 and visually in the second panel of Figure 5 (RQ 2). The figure clearly shows a modest but consistent increase in correlation when moving from a single call ($r = .71$) to a triple-call average ($r = .74$). However, the already strong performance of the single call, combined with the modest gain from averaging, indicates a high

Table 4. Correlation (Spearman's ρ) and 95% confidence intervals of prompting methods with human ratings using GPT-4o on the DS3 full dataset ($n = 421$).

Dimension	M1 vs Human	M2 vs Human	M3 vs Human
Solvability	.24 [−.01, .66]	.35 [−.01, .82]	−.01 [−.01, −.00]†
Line Elaboration	.64 [.54, .73]	.61 [.50, .70]	.56 [.45, .66]
Object Elaboration	.73 [.68, .78]	.71 [.66, .75]	.67 [.61, .72]
Aesthetics	.64 [.57, .70]	.58 [.51, .64]	.62 [.56, .68]
Humor & Surprise	.57 [.50, .63]	.45 [.37, .52]	.47 [.39, .54]
Title Creativity	.35 [.26, .44]	.34 [.25, .42]	.34 [.26, .43]
Originality	.53 [.46, .60]	.52 [.44, .59]	.48 [.40, .55]
Total Score	**.74 [.69, .79]**	**.71 [.65, .768]**	**.68 [.62, .74]**

Note: All correlations are significant at $p < .001$ (FDR-corrected), unless marked with † ($p > .05$). Values in brackets are 95% confidence intervals estimated via bootstrapping (1000 resamples).

Table 5. Correlation (Spearman's ρ) and 95% confidence intervals of prompting methods with human ratings using Gemini 2.5 Flash on the DS3 full dataset ($n = 421$).

Dimension	M1 vs Human	M2 vs Human	M3 vs Human
Solvability	.07 [−.03, .16]†	−.03 [−.12, .07]†	.21 [.12, .30]
Line Elaboration	.62 [.55, .67]	.59 [.52, .65]	.56 [.49, .62]
Object Elaboration	.59 [.52, .65]	.57 [.50, .63]	.51 [.44, .58]
Aesthetics	.57 [.50, .63]	.50 [.42, .57]	.48 [.40, .55]
Humor & Surprise	.36 [.27, .44]	.35 [.27, .44]	.33 [.24, .41]
Title Creativity	.17 [.07, .26]*	.18 [.09, .27]	.14 [.04, .23]*
Originality	.46 [.38, .53]	.42 [.34, .50]	.40 [.32, .48]
Total Score	**.57 [.51, .63]**	**.54 [.47, .61]**	**.48 [.41, .55]**

Note: All correlations are significant at $p < .001$ (FDR-corrected), unless marked otherwise: † $p > .05$; * $p < .01$. Values in brackets are 95% confidence intervals estimated via bootstrapping (1000 resamples).

Table 6. Performance of the M2 method: reliability interventions and model variations ($n = 421$).

Metric & Dimension	M2 Single Call (GPT-4o)	M2 Triple Avg (GPT-4o)	M2 Single Call (GPT-4o-mini)	M2 Single Call (GPT-4.1-nano)
Correlation (ρ) vs. Human	**.71 [.66, .76]**	**.74 [.69, .79]**	**.62 [.56, .68]**	**.41 [.33, .49]**
Mean (SD) of Scores	8.12 (2.59)	8.16 (2.50)	9.35 (8.92)	8.23 (5.70)
RMSE vs. Human	2.09	1.96	8.78	5.77
Correlations by Dimension (ρ)				
Solvability	.22 [−.01, .58]	.20 [−.01, .57]	−.02 [−.02, −.01]†	−.01 [−.01, −.001]†
Line Elaboration	.57 [.47, .67]	.64 [.54, .73]	.19 [.07, .30]	.19 [.10, .27]
Object Elaboration	.72 [.68, .77]	.75 [.70, .79]	.68 [.63, .73]	.60 [.53, .65]
Aesthetics	.61 [.55, .67]	.66 [.60, .71]	.35 [.26, .44]	.19 [.09, .28]
Humor & Surprise	.47 [.40, .54]	.53 [.46, .59]	.45 [.37, .52]	.18 [.09, .28]
Title Creativity	.33 [.25, .42]	.36 [.28, .44]	.33 [.24, .41]	.15 [.05, .24]*
Originality	.50 [.43, .57]	.57 [.50, .63]	.17 [.06, .25]*	.15 [.06, .24]*
Total Score	**.71 [.66, .76]**	**.74 [.69, .79]**	**.62 [.56, .68]**	**.41 [.33, .49]**

Note: This table compares the performance of different implementations of the M2 prompting method. The first section reports on the total creativity score, while the second section details correlations for each creativity dimension. Correlation (ρ) values are Spearman's rank-order correlations with human ratings; values in brackets are 95% confidence intervals (CIs) estimated via bootstrapping (1000 resamples). All correlations are significant at $p < .001$ (FDR-corrected), unless marked otherwise: † $p > .05$; * $p < .01$. The Mean (SD) row displays the mean and standard deviation of the scores generated by each method. RMSE is the Root Mean Square Error compared to human scores.

baseline level of reliability which offers a practical trade-off for applications where computational cost must (i.e., cost for three calls) be balanced against incremental gains in consistency.

Finally, we addressed RQ 3 by exploring the practical trade-offs between cost and quality across different GPT models. The results of this comparison are detailed in the final two columns of Table 6 and visualized in the third panel of Figure 5 (RQ 3). The figure illustrates a steep, stepwise degradation in performance as model size and cost decrease. The high-end GPT-4o set the standard ($r = .71$), followed by a notable decline for GPT-4o-mini ($r = .62$), and a substantial drop for GPT-4.1-nano ($r = .41$).

Discussion

This study examined the potential of LLMs, specifically GPT-4o and Gemini 2.5 Flash model families, to act as evaluators of creative game levels in a sandbox physics game called *Physics Playground*. This work contributes to the growing literature on AI-based creativity assessment. Overall, our findings demonstrate that rubric-guided LLMs can produce creativity ratings strongly aligned with human judgments when evaluating human-generated game levels in a multimodal learning game. These results advance recent work showing that GenAI can contribute to creativity assessment (Acar et al., 2024; Organisciak et al., 2023; Zhao et al., 2024) and extend it into ecologically valid, multimodal learning environments. The systematic process developed in this study can be extended to other content-creation learning games (e.g., *Minecraft*) that enable learners to generate artifacts within the game environment.

Beyond creativity assessment in learning games, this study provides a methodological roadmap for researchers who aim to use GenAI for educational assessment more broadly. By systematically testing prompting strategies, multimodal inputs, and model configurations, we illustrate how LLMs can be directed to produce responses that approximate human judgment. This structured approach

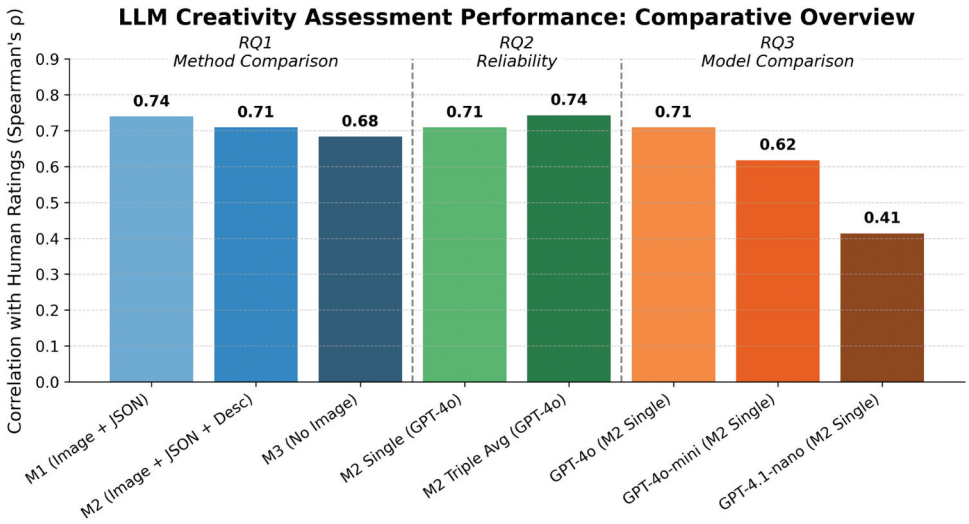


Figure 5. LLM creativity assessment performance: a comparative overview of prompting methods, reliability interventions, and model architectures.

can guide future studies seeking to adapt GenAI for other assessment domains. This iterative process ensures that validity and reliability are not left to chance but are achieved through careful design.

Validity and reliability of LLMs' creativity assessment

Findings provide strong evidence that rubric-guided prompting substantially enhances the validity of GPT-4o's creativity ratings. When LLMs were prompted without rubrics, their judgments resembled the consensual assessment technique (Amabile, 1982), which is valid for human experts but unsuitable for models lacking domain knowledge or intuitive expertise. Structured rubrics, previously validated in *Physics Playground* (Rahimi & Shute, 2021a), improved consistency and alignment with human raters, echoing findings from recent work on automated creativity assessment in verbal domains (Acar et al., 2024; Organisciak et al., 2023; Zhao et al., 2024). Relatedly, Haim et al. (2024) reported that LLMs diverge from human raters in how they weigh emotional and structural cues when judging textual creativity. This distinction highlights the importance of interpretability and guided evaluation frameworks. In our study, rubric-guided prompting served a similar function by making model reasoning more consistent and aligned with human evaluations in a multimodal game environment. Importantly, this study extends prior work by demonstrating that LLMs can assess both figural and verbal creativity within a dynamic learning environment – an advance over most research focused on static, text-based tasks.

Additional analyses showed that supplementing JSON and image inputs with LLM-generated textual descriptions did not improve validity, suggesting that multimodal input alone provides sufficient context and reduces computational cost. This finding suggests that the current JSON representation may lack important contextual or semantic details that images capture more effectively. Given that the game's JSON format was originally designed over a decade ago to serve the Unity engine – well before the advent of LLMs – it was not optimized for GenAI-based interpretation. Future work could augment these JSON files with richer metadata (e.g., creator-generated tags or descriptive labels) to better convey the conceptual and visual characteristics of each level, thereby enhancing model performance and interpretability.

The systematic performance gap between GPT-4o and Gemini (shown in Table 5) warrants consideration. While both models exhibited similar dimension-level patterns, Gemini's correlations were consistently 0.17–0.20 points lower across all input configurations. Several factors may contribute to this difference. First, architectural differences in multimodal fusion may affect how each model integrates visual and structured information. GPT-4's training emphasized tight integration of vision and language representations, which may be particularly advantageous for tasks requiring simultaneous interpretation of images and structured data like JSON. Second, differences in training data composition and objectives may influence how well each model aligns with human creativity judgments in game-based contexts. Third, our prompting approach, while designed to be model-agnostic, may have been inadvertently optimized for GPT-like architectures through iterative refinement on GPT-4o during the preliminary phases. Despite this performance gap, the finding that both models showed consistent method-level patterns (M1 > M2 >M3) across all creativity dimensions suggests that the core methodological principles, rubric-guided prompting with multimodal input, generalize beyond specific model implementations. This cross-architecture validation strengthens confidence that observed effects reflect genuine assessment validity rather than model-specific artifacts.

Reliability tests indicated that a single call to GPT-4o was nearly as effective as averaging across three calls, highlighting the scalability of this approach. Finally, comparisons across model families confirmed that larger LLMs consistently outperformed smaller variants, reinforcing prior evidence that high-capacity models better capture nuanced dimensions of creativity (Bellemare-Pepin et al., 2024; Zhao et al., 2024). Taken together, the optimal configuration for LLM-based creativity assessment involves a single call to a large model, guided by a validated rubric, with JSON and image input.

Future directions and ethical considerations

Future work should focus on methodological, technical, and educational advances. Methodologically, originality assessment remains constrained by the absence of historical comparison. Incorporating retrieval-augmented generation (Leng et al., 2024) would allow models to situate new levels against prior designs, strengthening judgments of statistical rareness (Runco & Acar, 2012). Technical advances should target refining rubric criteria for title creativity, which showed systematically lower correlations across all models and methods. Educationally, these advances could enable more personalized gameplay, aligning with efforts in adaptive learning to tailor tasks to students' interests and prior knowledge (Holmes et al., 2018; Pane et al., 2015; Rahimi & Shute, 2025; Walkington, 2013). Such personalization has the potential to boost interest and engagement, both of which are strong

predictors of learning (Ainley, 2012; Guzey & Li, 2023; Hidi & Renninger, 2006).

Another promising direction is the development of AI coaches that not only assess but also support creativity in real time. Building on the approach demonstrated here, GenAI could serve as an interactive creativity coach that evaluates learners' designs, provides tailored feedback, and offers scaffolds to encourage divergent thinking. Such AI-powered coaching systems could expand the reach of creativity support in classrooms and help students build confidence and skill in both creative problem solving and disciplinary learning.

Additionally, the ethical use of GenAI in creativity assessment warrants careful consideration, particularly in educational contexts involving students. Recent reviews emphasize that while AI can enhance personalization and assessment, it can also perpetuate bias, reduce transparency, and risk inequitable learning outcomes if not carefully designed (Chinta et al., 2024; García-López & Trujillo-Liñán, 2025). Algorithmic bias remains a central concern, as models trained on large datasets may encode cultural or linguistic patterns that undervalue diverse forms of creative expression (Rahimi et al., in press; Dieterle et al., 2024). Transparency and accountability are equally critical, ensuring that educators and learners can interpret AI-generated evaluations and understand their limitations (Cheong, 2024). Moreover, the balance between fairness and predictive accuracy presents an ongoing challenge for applying AI ethically in educational assessment (Chinta et al., 2024). AI-based small-scale creativity assessments could serve a valuable formative role when implemented with human oversight and transparency. This design may ensure that automated creativity assessments complement rather than replace human judgment. Future applications of large-scale GenAI in creativity assessment should rigorously examine fairness and related ethical dimensions to ensure the full psychometric quality of such assessments. Periodic calibration against human ratings may be necessary to detect and correct deviations from fair and ethical interpretations.

Lastly, future research could investigate how GenAI can help develop endlessly adaptive learning games that personalize challenge and content to learners' needs and interests. Prior work shows that GenAI can support machine creativity (Bellemare-Pepin et al., 2024; Haase & Hanel, 2023) and generate novel artifacts (Boden, 2004; Ma et al., 2023), though human work still can surpass AI in creativity (e.g., Chakrabarty et al., 2024). LLMs have also been used to generate game levels across asset-based (i.e., putting together a new game level using existing artifacts based on a prompt), non-asset-based (i.e., creating all aspects of a game level from scratch), and mixed approaches. For example, Sokoban and VGDL levels (Hu et al., 2024; Todd et al., 2023), asset-driven scene creation (Hu et al., 2024; Kumaran et al., 2024), narrative-to-world mappings (Nasir et al., 2024), novel puzzles (Todd et al., 2024), and existing games such as *Minecraft* (Huang et al., 2025). Building on these advances, future educational games could use GenAI to generate personalized, continuously evolving games that adapt to learner preferences, learning trajectories, and motivational needs, thereby enhancing engagement, agency, and the overall learning experience.

Implications

The findings have implications for creativity research, assessment, and educational practice. For theory, they contribute to ongoing debates about whether AI can engage in genuine creativity (Lockhart, 2025; Runco, 2023), while recognizing that current outputs reflect "artificial creativity" (Runco, 2023) that complements rather than replaces human originality. For assessment, rubric-guided prompting illustrates how LLMs can act as scalable "third raters," providing valid and reliable judgments of multimodal creative products (Rahimi et al., 2024; Shute & Wang, 2016). For education, embedding these assessments in *Physics Playground* demonstrates how AI can unobtrusively support creativity development in real time, reinforcing calls for creativity as a 21st-century competency (Craft, 2010; Glaveanu et al., 2020; UNESCO, 2023). The work also advances AI literacy by showing how human-designed scaffolds (e.g., rubrics) are essential for steering AI behavior, offering a model of responsible human – AI collaboration (Holstein & Doroudi, 2022).

Limitations

Several methodological and practical considerations constrain the scope of our findings. This study focused on visual-spatial creativity in game level design, a domain characterized by multimodal artifacts combining structured data with visual representations. The extent to which rubric-guided LLM assessment generalizes to other creative domains, such as purely verbal tasks or divergent thinking assessments, remains an empirical question requiring domain-specific validation. This trade-off between ecological validity in authentic learning environments and generalizability across creative domains reflects a fundamental challenge in creativity research that should be addressed.

Rapid evolution in LLM architectures introduces uncertainty regarding long-term reproducibility. While our cross-architecture validation (GPT-4o and Gemini 2.5 Flash) demonstrates that methodological principles

hold across distinct model families, periodic revalidation with emerging models will be necessary as older systems become legacy infrastructure. Importantly, our objective was not to benchmark state-of-the-art performance but to examine whether rubric-guided multimodal prompting generalizes across architecturally distinct systems. As newer versions of GPT and Gemini models are released, future work should reassess these findings to ensure continued robustness. The substantial performance gap between high-capacity (GPT-4o) and lower-cost models (GPT-4o-mini, GPT-4.1-nano) also highlights ongoing tensions between assessment quality and computational accessibility in educational settings.

Inference configuration represents another boundary condition. Although fully deterministic inference (e.g., temperature = 0) could be recommended for automated scoring to minimize stochastic variation, we employed moderate temperature settings with strict JSON schema enforcement to reflect realistic API-based deployment conditions. Empirically, single-call scoring demonstrated strong agreement with human ratings, and averaging across multiple calls yielded only modest gains, suggesting that rubric-guided prompting and structured output constraints substantially stabilize scoring behavior even under non-zero temperature conditions. Nevertheless, future research should directly compare deterministic (temperature = 0) and ensemble-based approaches to more precisely quantify trade-offs between inference stability, computational cost, and validity in multimodal creativity assessment tasks.

Title creativity presented persistent assessment challenges across all models and methods, with correlations substantially lower than other dimensions. This likely reflects the inherent difficulty of evaluating brief verbal creativity that relies heavily on cultural context, linguistic play, and subjective humor. Whether this represents a fundamental limitation of current LLMs or suggests the need for alternative assessment approaches warrants further investigation.

Finally, while our study provides convergent validity evidence, questions of algorithmic bias, fairness across diverse student populations, and transparency in AI-driven educational assessment remain critical concerns requiring systematic empirical examination (Holmes & Porayska-Pomsta, 2023).

Conclusion

This study provides evidence that large language models, when guided by a validated rubric, can produce creativity ratings strongly aligned with human judgments in a non-asset-based educational game environment. By demonstrating that GPT-4o can assess both figural and verbal creativity in *Physics Playground* using multimodal input (JSON and images), this work extends prior research on AI-based creativity assessment beyond text-based domains into ecologically valid, dynamic contexts. Importantly, these findings highlight the value of rubric-guided prompting for ensuring validity and reliability, while also showing the promise and limitations of LLMs as creative agents.

Looking ahead, further advances in originality detection and rubric refinement will be essential to realize the full potential of LLM-based creativity assessment systems. Such developments could enable more personalized, engaging gameplay, fostering both creativity and disciplinary learning in STEM contexts. Beyond technical contributions, this work also aligns with broader educational goals: embedding unobtrusive creativity assessment within evidence-centered design frameworks (Almond et al., 2015; Rahimi & Shute, 2024; Shute, 2011) and support creativity as a 21st-century competency, and advancing AI literacy by modeling responsible human – AI collaboration. With these improvements, LLMs can serve not as replacements for human judgment, but as scalable partners that help measure and nurture creativity while simultaneously supporting learning.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Seyedahmad Rahimi http://orcid.org/0000-0001-9266-758X
Hongming Li http://orcid.org/0000-0002-6024-677X
Salah Esmaeiligoujar http://orcid.org/0000-0002-1899-8660
Deniz Ercan http://orcid.org/0000-0002-3436-6444
Anthony Botelho http://orcid.org/0000-0002-7373-4959

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT to proofread and improve the writing's clarity. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Acar, S. (2025). Creativity assessment, research, and practice in the age of artificial intelligence. *Creativity Research*

Journal, *37*(2), 181–187. https://doi.org/10.1080/10400419.2023.2271749

Acar, S., Dumas, D., Organisciak, P., & Berthiaume, K. (2024). Measuring original thinking in elementary school: Development and validation of a computational psychometric approach. *Journal of Educational Psychology*, *116*(6), 953–981. https://psycnet.apa.org/doi/10.1037/edu0000844

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv: 2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

Ainley, M. (2012). Students' interest and engagement in classroom activities. In S. L. Christenson, et al. (Ed.), *Handbook of research on student engagement* (pp. 283–302). Springer.

Almond, R. G., Mislevy, R. J., Steinberg, L. S., Yan, D., & Williamson, D. M. (2015). *Bayesian networks in educational assessment*. Springer.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality & Social Psychology*, *43*(5), 997. https://doi.org/10.1037/0022-3514.43.5.997

Amabile, T. M. (2018). *Creativity in context: Update to the social psychology of creativity*. Routledge.

Anantrasirichai, N., Zhang, F., & Bull, D. (2025). Artificial intelligence in creative industries: Advances prior to 2025. *arXiv preprint arXiv: 2501.02725*. https://arxiv.org/abs/2501.02725

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal*, *34*(3), 245–260. https://doi.org/10.1080/10400419.2022.2025720

Beghetto, R. A., & Karwowski, M. (2023). Creative self-beliefs: From creative potential to creative action. In R. J. Sternberg & J. C. Kaufman (Eds.), *Handbook of organizational creativity* (2nd ed. pp., 179–193). Academic Press. https://doi.org/10.1016/B978-0-323-91840-4.00010-4

Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., Bengio, Y., & Jerbi, K. (2024). Divergent creativity in humans and large language models. *arXiv preprint arXiv: 2405.13012*. https://doi.org/10.48550/arXiv.2405.13012.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms* (2nd ed.). Routledge. https://doi.org/10.4324/9780203508527

Bowman, S. R. (2023). Eight things to know about large language models. *arXiv. arXiv preprint arXiv: 2304.00612*. https://doi.org/10.48550/arXiv.2304.00612

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Cai, W., & Gao, M. (2025). Beyond hallucination: Generative AI as a catalyst for human creativity and cognitive evolution. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, *2*(1), 36–42. https://doi.org/10.62762/TETAI.2025.657559

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C. S. (2024, May). Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the CHI conference on human factors in computing systems* (pp. 1–34). https://doi.org/10.1145/3613904.3642731

Chaudhry, M. A., & Kazim, E. (2022). Artificial intelligence in education (AIEd): A high-level academic and industry note 2021. *AI and Ethics*, *2*(1), 157–165. https://doi.org/10.1007/s43681-021-00074-z

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv*: 2107.03374.

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, *6*, 1421273. https://doi.org/10.3389/fhumd.2024.1421273

Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. *arXiv preprint arXiv: 2407.18745*.

Craft, A. (2010). *Creativity and education futures: Learning in a digital age*. Trentham.

Cropley, A., & Cropley, D. (2011). Creativity and lawbreaking. *Creativity Research Journal*, *23*(4), 313–320. https://doi.org/10.1080/10400419.2011.621817

DiBattista, A., Grayling, S., Hasselaar, E., Leopold, T., Li, R., Rayner, M., & Zahidi, S. (2023, May). *Future of jobs report 2023*. *In World Economic Forum*. https://www.weforum.org/publications/the-future-of-jobs-report-2023/

Dieterle, E., De de, C., & Walker, M. (2024). The cyclical ethical effects of using artificial intelligence in education. *AI & Society*, *39*(2), 633–643. https://doi.org/10.1007/s00146-022-01497-w

DiStefano, P. V., Patterson, J. D., & Beaty, R. E. (2025). Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, *37*(4), 555–569. https://doi.org/10.1080/10400419.2024.2326343

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., & Welty, C. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, *31*(3), 59–79. https://doi.org/10.1609/aimag.v31i3.2303

García-López, A., & Trujillo-Liñán, J. (2025). Ethical and regulatory challenges of generative AI in education: A systematic review. *Frontiers in Education*, *10*(1565938). https://doi.org/10.3389/feduc.2025.1565938

Gee, J. P. (2005). Learning by design: Good video games as learning machines. *e-Learning & Digital Media*, *2*(1), 5–16. https://doi.org/10.2304/elea.2005.2.1.5

Glaveanu, V. P., Hanchett Hanson, M., Baer, J., Barbot, B., Clapp, E. P., Corazza, G. E., & Sternberg, R. J. (2020). Advancing creativity theory and research: A socio-cultural manifesto. *Journal of Creative Behavior*, *54*(3), 741–745. https://doi.org/10.1002/jocb.395

Goecke, B., DiStefano, P. V., Aschauer, W., Haim, K., Beaty, R., & Forthmann, B. (2024). Automated scoring of scientific creativity in German. *Journal of Creative Behavior*, *58*(3), 321–327. https://doi.org/10.1002/jocb.658

Gu, X., & Ericson, B. J. (2025). AI literacy in K-12 and higher education in the wake of generative AI: An integrative review. In L. Porter B. Morrison N. Brown C. S. Montero (Eds.). In *Proceedings of the 2025 ACM Conference on International Computing Education Research* (Vol. 1. pp. 125–140).

Guzey, S. S., & Li, W. (2023). Engagement and science achievement in the context of integrated STEM education: A longitudinal study. *Journal of Science Education and Technology*, *32*(2), 168–180. https://doi.org/10.1007/s10956-022-10023-y

Haase, J., & Hanel, P. H. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *Journal of Creativity*, *33*(3), 100066. https://doi.org/10.0066.10.1016/j.yjoc.2023.100066

Haase, J., & Pokutta, S. (2024). Human-AI co-creativity: Exploring synergies across levels of creative collaboration. *arXiv preprint arXiv: 2411.12527*. https://arxiv.org/abs/2411.12527

Hadas, E., & Hershkovitz, A. (2025). Assessing creativity across multi-step intervention using generative AI models. *Journal of Learning Analytics*, *12*(1), 91–109. https://doi.org/10.18608/jla.2025.8571

Haim, E., Fischer, N., Citraro, S., Rossetti, G., & Stella, M. (2024). Forma mentis networks predict creativity ratings of short texts via interpretable artificial intelligence in human and GPT-simulated raters. *arXiv preprint arXiv: 2412.00530*.

Heigl, R. (2025). Generative artificial intelligence in creative contexts: A systematic review and future research agenda. *Management Review Quarterly*, 1–38. https://doi.org/10.1007/s11301-025-00494-9

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111–127. https://doi.org/10.1207/s15326985ep4102_4

Holmes, W., Bialik, M., & Fadel, C. (2018). *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign.

Holmes, W., & Porayska-Pomsta, K. (Eds.). (2023). *The ethics of artificial intelligence in education: Practices, challenges, and debates*. Routledge. https://doi.org/10.4324/9780429329067

Holstein, K., & Doroudi, S. (2022). Equity and artificial intelligence in education. In W. Holmes, K. Porayska-Pomsta (Eds.). *The ethics of artificial intelligence in education* (pp. 151–173). Routledge.

Hu, C., Zhao, Y., & Liu, J. (2024, August). Game generation via large language models. In A. Dockhorn D. Loiacono E. Mekler P. Burelli M. Jiwatode (Eds.). *2024 IEEE conference on games (CoG)* (pp. 1–4). IEEE. https://doi.org/10.48550/arXiv.2404.08706

Huang, S., Nasir, M. U., James, S., & Togelius, J. (2025, March). Word2minecraft: Generating 3d game levels through large language models. *arXiv*. https://doi.org/10.48550/arXiv.2503.16536

Irmak, N. (2024). Artifacts without authors: Generative artificial intelligence and the question of authorship. *Metaphysics*, *7*(1), 1–15. https://doi.org/10.5334/met.160

JSON (JavaScript Object Notation). (2023). *Introducing json*. https://www.json.org/json-en.html

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four C model of creativity. *Review of General Psychology*, *13*(1), 1–12. https://doi.org/10.1037/a0013688

Kaufman, J. C., & Sternberg, R. J. (Eds.). (2010). *The Cambridge handbook of creativity*. Cambridge University Press.

Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, *13*(1), 15739. https://doi.org/10.1038/s41598-023-40858-3

Kumaran, V., Rowe, J., & Lester, J. (2024, November). Narrativegenie: Generating narrative beats and dynamic storytelling with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *20*(1), 76–86. https://doi.org/10.1609/aiide.v20i1.31868

Kwon, H., & Lee, Y. (2025). A meta-analysis of STEM project-based learning on creativity. STEM Education, 5(2), 275–290.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In D. Forsyth V. Štruc (Eds.). *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13872–13882). https://doi.org/10.1109/CVPR52733.2024.01316

Li, J., Yuan, Y., & Zhang, Z. (2024). Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv*. https://arxiv.org/abs/2403.10446

Lockhart, J. (2025). Creativity in the age of AI: The human condition and the limits of machine imagination. *AI & Society*, *40*(1), 123–135. https://doi.org/10.1007/s41809-024-00158-2

Luchini, S. A., Pronchick, J., Ceh, S., Kaufman, J. C., Johnson, D., Rafner, J., & Beaty, R. E. (2025). *The roles of idea generation and elaboration in human-AI collaborative creativity [Preprint]. PsyArXiv*. https://doi.org/10.31234/osf.io/xm2f5_v2

Luthra, P. (2025). TraitSpaces: Towards interpretable visual creativity for human-AI co-creation. *arXiv preprint arXiv: 2509.24326*.

Ma, K., Grandi, D., McComb, C., & Goucher-Lambert, K. (2023, August). Conceptual design generation using large language models. In S. Goyal A. Müller F. Ahmed B. Morkos (Eds.). *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 87349, pp. V006T06A021). American Society of Mechanical Engineers.

Mondorf, L., & Plank, B. (2024). Beyond accuracy: Evaluating the reasoning behavior of large language models. *arXiv preprint arXiv: 2404.01869*. https://arxiv.org/abs/2404.01869

Nasir, M. U., James, S., & Togelius, J. (2024). Word2world: Generating stories and worlds through large language models. *arXiv preprint arXiv: 2405.06686*.

Neves, M. C. (2024). *Prompt engineering vs. RAG vs. fine-tuning: What do you need?* TensorOps. https://www.tensorops.ai/post/prompt-eng-vs-rag-vs-fine-tuning-what-do-you-need

Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2023). Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint* arXiv: 2401.00396.

OECD. (2024). *Global trends in government innovation, 2024: Fostering human-centred public services*, OECD Public Governance Reviews, OECD Publishing. https://doi.org/10.1787/c1bc19c3-en

Organisciak, P., Acar, S., Dumas, D., & Berthiaume, K. (2023). Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, *49*, 101356. https://doi.org/10.1016/j.tsc.2023.101356

Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2015). *Continued progress: Promising evidence on personalized learning*. RAND Corporation.

Plucker, J. A., Beghetto, R. A., & Dow, G. T. (2004). Why isn't creativity more important to educational psychologists? Potentials, pitfalls, and future directions in creativity research. *Educational Psychologist*, *39*(2), 83–96. https://doi.org/10.1207/s15326985ep3902_1

Qian, M., & Clark, K. R. (2016). Game-based learning and 21st century skills: A review of recent research. *Computers in Human Behavior*, *63*, 50–58. https://doi.org/10.1016/j.chb.2016.05.023

Rafner, J., Biskjær, M. M., Zana, B., Langsford, S., Bergenholtz, C., Rahimi, S., Carugati, A., Noy, L., & Sherson, J. (2022). Digital games for creativity assessment: Strengths, weaknesses and opportunities. Creativity Research Journal, 34 (1), 28–54. https://doi.org/10.1080/10400419.2021.1971447

Rahimi, S. (2020). Inspire, Instruct, or Both? Game-based assessment and support of creativity. (Doctoral Dissertation). Florida State University

Rahimi, S. (2023). Going Beyond the Brick: Assessing and Supporting Creativity Using AI-Powered Digital Games. Creativity Research Journal, 1–9. http://dx.doi.org/10.1080/10400419.2023.2241779

Rahimi, S., Almond, R., & Shute, V. J. (2023). Stealth assessment's technical architecture. In M. P. McCreery, & S. K., Krach (Eds.), Games as stealth assessments, (pp. 61–80). Hershey, PA: IGI Global.

Rahimi, S., Babaee, M., Esmaeiligoujar, S., & Dede, C. (in press). Creativity and generative AI. In R. A. Beghetto (Ed.), The Oxford handbook of AI and creativity in education. Oxford University Press.

Rahimi, S., Dede, C., Esmaeiligoujar, S., & Babaee, M. (2026). Augmenting human creativity with responsible and ethical use of generative AI. In M. J. Worwood & J. C. Kaufman (Eds.), Generative artificial intelligence and creativity (Explorations in Creativity Research, pp. 87–99). Academic Press. https://doi.org/10.1016/B978-0-443-34073-4.00010-1

Rahimi, S., & Shute, V. J. (2021a). First inspire, then instruct to improve students' creativity. Computers & Education, 174, 104312.ht tps://psycnet.apa.org/d oi/1 0.1016/j.compedu.2021.104312

Rahimi, S., & Shute, V. J. (2021b). The effects of video games on creativity: A systematic review. In S. W. Russ, J. D. Hoffmann, & J. C. Kaufman (Eds.), Handbook of lifespan development of creativity (pp. 368–392). Cambridge University Press. https://doi.org/10.1017/9781108755726.021

Rahimi, S., & Shute, V. J. (2024). Stealth assessment: a theoretically grounded and psychometrically sound method to assess, support, and investigate learning in technology-rich environments. Educational technology research and development, 1–25. https://doi.org/10.1007/s11423-023-10232-1

Rahimi, S., & Shute, V. J. (2025). Personalized learning in educational games using stealth assessment. In M. L. Bernacki, C. Walkington, A. Emery, & L. Zhang (Eds.), Handbook of Personalized Learning (Chapter 5). (Eds.), Handbook of Personalized Learning (Chapter 5). New York, NY: Routledge. https://doi.org/10.4324/9781032719467-7

Rahimi, S., Smith, J. B., Truesdell, E. J., Vinay, A., Boyer, K. E., Magerko, B., & Mcklin, T. (2024). An automated, unobtrusive, formative assessment of creativity in a computer science and music remixing learning environment. Psychology of Aesthetics, Creativity, and the Arts. https://doi.org/10.1037/aca0000683

Runco, M. A. (2023). AI can only produce artificial creativity. *Journal of Creativity*, *33*(3), 100063. https://doi.org/10.1016/j.yjoc.2023.100063

Runco, M. A., & Acar, S. (2012). Divergent thinking as an indicator of creative potential. *Creativity Research Journal*, *24*(1), 66–75. https://psycnet.apa.org/doi/10.1080/10400419.2012.652929

Seli, P., Ford, T. E., Kuan, M., Leung, E., & Hopp, H. (2025). Beyond the brush: Comparing human and AI creativity in the domain of digital art. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication. https://doi.org/10.1037/aca0000743

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv: 2104.07567*.

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, *55*(2), 503–524.

Shute, V. J., Almond, R. G., & Rahimi, S. (2019). Physics Playground (version 1.3)[computer software]. Tallahassee, FL.

Shute, V. J., & Rahimi, S. (2021). Stealth assessment of creativity in a physics video game. Computers in Human Behavior, 116, 106647. https://doi.org/10.1016/j.chb.2020.106647

Shute, V. J., & Wang, L. (2016). Assessing and supporting hard-to-measure constructs. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and application* (pp. 535–562). John Wiley & Sons, Inc.

Singh, J., Magazine, R., Pandya, Y., & Nambi, A. (2025). Agentic reasoning and tool integration for LLMs via reinforcement learning. https://www.microsoft.com/en-us/research/wp-content/uploads/2025/04/AgenticReasoning.pdf

Sternberg, R. J., & Lubart, T. I. (1996). Investing in creativity. *The American Psychologist*, *51*(7), 677–688. https://doi.org/10.1037/0003-066X.51.7.677

Syamra, F. M., & Suryadi, A. (2025). Comparing Project-Based Learning and Integrated STEM Approach in Enhancing Scientific Creativity in Renewable Energy Education. Pedagogical Research, 10(3).

Todd, G., Earle, S., Nasir, M. U., Green, M. C., & Togelius, J. (2023, April). Level generation through large language models. In P. Lopes F. Luz A. Liapis H. Engström (Eds.). In *Proceedings of the 18th international conference on the foundations of digital games*: (pp. 1–8). New York: ACM. https://doi.org/10.48550/arXiv.2302.05817

UNESCO. (2023). *Global education monitoring report, 2023: Technology in education: A tool on whose terms?* https://doi.org/10.54676/UZQV8501

Urmeneta, A., & Romero, M. (2025, August). AI as a creative partner: A prisma review of AI's role in supporting creativity in education. *Frontiers in Education*, *10*, 1602151. https://doi.org/10.3389/feduc.2025.1602151

Walkington, C. (2013). Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932–945. https://doi.org/10.1037/a0031882

Wingström, R., Hautala, J., & Lundman, R. (2024). Redefining creativity in the era of AI? Perspectives of computer scientists and new media artists. *Creativity*

Research Journal, *36*(2), 177–193. https://doi.org/10.1080/10400419.2022.2107850

World Economic Forum. (2023, April 30). *The future of jobs report 2023*. http://www3.weforum.org/docs/WEF_Future_of_Jobs_2023.pdf

World Economic Forum. (2025, January 7). *The future of jobs report 2025*. https://www.weforum.org/publications/the-future-of-jobs-report-2025/in-full/

Zhai, X. (2024). Transforming teachers' roles and agencies in the era of generative AI: Perceptions, acceptance, knowledge, and practices. *Journal of Science Education and Technology*, 1–11. https://doi.org/10.1007/s10956-024-10174-0

Zhang, C., & Xu, S. (2025). *Aesthetic experience and educational value in co-creating art with generative AI: Evidence from a survey of young learners*. [Preprint]. arXiv.

Zhao, Y., Zhang, R., Li, W., Huang, D., Guo, J., Peng, S., & Chen, Y. (2024). Assessing and understanding creativity in large language models. Machine Intelligence Research, 22, 417–436. https://doi.org/10.1007/s11633-025-1546-4