Technovation 155 (2026) 103571

Contents lists available at ScienceDirect

Technovation

journal homepage: www.elsevier.com/locate/technovation

Evaluating creative work with artificial intelligence: Evidence from constrained innovation tasks☆

Valerio Fedele Addis [a], Giuseppe Attanasi [a,b,*], Giovanni Di Bartolomeo [a,c], Michele Mariella [a], Valentina Peruzzi [a]

[a] Department of Economics and Law, Sapienza University of Rome, Italy
[b] Institute of Entrepreneurship and Innovation, Corvinus University of Budapest, Hungary
[c] Corvinus Institute for Advanced Studies (CIAS), Corvinus University of Budapest, Hungary

ARTICLE INFO

JEL classification:
O31
D83
M14
C91

Keywords:
Artificial intelligence
Creativity evaluation
Constrained creativity tasks
Consensual assessment technique
Cultural and creative industry professionals
Innovation-like tasks

ABSTRACT

We study whether a large language model can reliably evaluate human creativity in constrained, innovation-like tasks. Using expert-generated creative outputs from a validated experiment with workers in cultural and creative industries, we embed ChatGPT as an evaluator and benchmark its assessments against expert human judgments obtained through the Consensual Assessment Technique. Study 1 supports AI reliability by showing that AI-based creativity evaluations exhibit internal consistency comparable to that of expert judges across repeated and independent runs, even under conservative scenarios. Replacing a human judge with an AI evaluator does not reduce inter-rater reliability across drawing, mathematical, and verbal tasks. Beyond reliability, AI evaluations display three additional features that are difficult to achieve with human-only panels: lower evaluative variability, systematically higher scores consistent with a potentially more inclusive evaluative stance, and task-independence of evaluative standards. Study 2 further supports task-independence by showing that AI evaluations are structured along fluency, flexibility, originality, and elaboration, with dimension weights that adapt to task-specific constraints.

1. Introduction

Generative artificial intelligence is rapidly reshaping how organizations produce, evaluate, and select creative ideas. While much of the existing debate has focused on AI as a tool for content generation, an equally consequential transformation concerns the role of AI in evaluating, selecting, and legitimizing creativity. In innovation processes, evaluation is not a neutral step: it determines which ideas are funded, which projects are pursued, and which creators receive recognition and rewards. As firms increasingly rely on algorithmic systems to support decision-making in innovation (Roberts and Candi, 2024) and hiring (Abrardi et al., 2022; Li et al., 2026), understanding how AI evaluates creativity becomes critical for the governance of creative work and innovation processes.

Despite this growing relevance, existing research has mostly examined AI as a creator and humans as evaluators of AI outputs (e.g., Charness and Grieco, 2026; Magni et al., 2024; Bellemare-Pepin et al., 2026; Arora et al., 2025), almost overlooking the reverse configuration: AI as an evaluator of human creativity. This gap is particularly significant for organizations, where scalable evaluation technologies may influence idea screening, resource allocation, and the selection of innovative projects under uncertainty. If AI systems increasingly act as gatekeepers in innovation processes, their internal logic of creativity assessment may shape not only which ideas succeed, but also which forms of creativity are systematically rewarded.

This paper addresses this gap by investigating whether a large language model can function as a reliable and informative evaluator of human creativity in constrained, innovation-like environments. Leveraging expert-generated creative outputs from a validated experimental design, we embed an AI evaluator (ChatGPT) into the creativity assessment process through a standardized and replicable prompting protocol. A distinctive feature of our approach is that we explicitly account for the stochastic nature of large language model outputs.

☆ This research has been funded by the Italian Ministry of Universities and Research under Grant PRIN 2022 n. 20229LRAHK (funded by the European Union - Next Generation EU, Mission 4 Component 1, CUP B53D23012680006), the Sapienza University under Ateneo Grants 2023 for project n. RG123188B4CEE028 and the D34Health through the PNC-Spoke3 program research project "Wearable technologies, sensors, and biomarkers for care through digital twin approaches", under Grant B53C22006120001.

* Correspondence to: Department of Economics and Law, Sapienza University of Rome, Via del Castro Laurenziano, 9 - 00161 Rome, Italy
E-mail address: giuseppe.attanasi@uniroma1.it (G. Attanasi).

https://doi.org/10.1016/j.technovation.2026.103571
Received 31 January 2026; Received in revised form 18 April 2026; Accepted 26 April 2026
Available online 14 May 2026

Rather than treating AI evaluation as a single judgment, we study a distribution of independent AI evaluations and assess whether reliability holds across different realizations of the model. This allows us to evaluate not only average performance, but also the stability and portability of AI-based creativity assessment under repeated use.

Our analysis proceeds in two complementary studies. Study 1 examines whether AI-generated creativity evaluations display levels of reliability comparable to those of expert human judges when embedded in a Consensual Assessment Technique (CAT) framework. We treat each independent AI run as a potential evaluator and study the distribution of inter-rater reliability obtained when the AI is added to a panel of four human judges. Adopting a deliberately conservative perspective, we focus on two polar realizations of the AI evaluator: the iteration associated with the highest internal consistency and the one associated with the lowest internal consistency across repeated runs. This design allows us to assess whether AI-based evaluation remains reliable even in the least favorable realizations of the model's stochastic output. Study 2 complements these findings by examining the internal structure of AI creativity judgments. By decomposing overall creativity scores into multiple dimensions inspired by the Torrance (1974) framework – fluency, flexibility, originality, and elaboration – and by comparing AI evaluations across task domains, we investigate whether AI relies on a coherent and theoretically grounded evaluative structure, and whether this structure adapts to differences in task constraints. By comparing AI's overall creativity scores across Study 1 and Study 2, we also verify that AI evaluations are essentially invariant to the explicit association of specific criteria with creativity assessment.

A central feature of our approach concerns the nature of the creative tasks used to study creativity and innovation. Much of the empirical literature relies on unconstrained or loosely structured tasks, such as free-form ideation or divergent thinking exercises (Bellemare-Pepin et al., 2026; Arora et al., 2025). While informative about individual creative potential, these tasks differ markedly from the environments in which organizational innovation typically unfolds. In practice, innovation takes place under technical, organizational, and economic constraints that define objectives, delimit admissible solutions, and shape both creative production and its evaluation (Hewitt-Dundas, 2006; Hoegl et al., 2008; Hottenrott and Peters, 2012; Murro and Peruzzi, 2026).

Experimental research on creativity has increasingly emphasized the importance of task structure in capturing these features. In particular, closed creativity tasks, characterized by clearly specified goals and binding constraints on allowable outputs, have been proposed as laboratory analogues of innovation under constraints. A distinctive contribution in this tradition is provided by Charness and Grieco (2019, 2023), who design a set of closed tasks of fundamentally different nature (drawing, mathematical, and verbal), that nonetheless generate comparable outcomes in terms of materials (pen and paper) and format (a single A4 sheet).[1] This comparability allows creativity to be studied across heterogeneous domains while holding constant the physical environment and the nature of the output, and it also facilitates standardized evaluation by external judges, including algorithmic evaluators.

These tasks are deliberately simple and stylized. While they abstract from many real-world features of innovation processes, this abstraction is precisely what makes them analytically powerful. By isolating specific creative mechanisms within tightly constrained environments, closed tasks capture distinct modes of innovation in a generalizable and portable way. The *drawing task*, which requires the recombination of predefined elements into a novel visual artifact, resembles *product-oriented innovation*, where creativity manifests through design choices under material and aesthetic constraints. The *mathematical task*, centered on transforming given inputs into a target outcome through admissible operations, mirrors *process innovation*, emphasizing rule-based problem-solving within strict technical constraints. This structure has concrete real-world analogues. In algorithm and procedure design, the objective is fixed but multiple valid solution paths exist, differing in elegance, elaboration, and novelty of approach. In patent design-around strategies, firms must achieve the same functional outcome through a different and well-articulated chain of technical transformations, with creativity evaluated explicitly on the basis of novelty and inventive step. In both cases, as in the mathematical task, the endpoint is fixed while the creative margin lies entirely in the construction of the solution path.[2] The *verbal task*, which involves constructing a coherent narrative from a fixed set of elements, reflects *symbolic or meaning-based innovation*, where value is created through interpretation, framing, and recombination.

An additional and often overlooked dimension of research on creativity and innovation evaluation concerns the composition of subject pools. The overwhelming majority of experimental studies rely on university students as experimental subjects (Arechar et al., 2018), a pattern that also characterizes experimental research on creativity in the psychology, management and economics fields (Attanasi et al., 2021). As a result, empirical evidence on creativity evaluation is largely based on populations with often limited creative experience. Only a small number of studies involve non-student participants (e.g., Ariely et al., 2009), and even fewer focus explicitly on creativity experts.[3] Attanasi et al. (2026) provide one of the few experimental economics studies that directly compare experts and non-experts performing the same closed creativity tasks. Their design involves both students and creativity experts completing the three comparable tasks used in Charness and Grieco (2019, 2023) — drawing, mathematical, and verbal. Importantly, among their 120 expert participants, a subset (approximately 30 subjects) consists of professionals actively operating in creative industries. The creative outputs produced by these professionals (approximately 10 per task domain) constitute the empirical foundation of the present analysis.

More precisely, we take as benchmark the double-blind evaluations conducted by the four external judges in Attanasi et al. (2026) of the creative artifacts produced by creative-industry professionals. Building on them, we conduct an AI-based creativity evaluation designed to mirror, as closely as possible, the informational and procedural conditions faced by human judges. For each task domain, the model is provided with the original task instructions and the corresponding set of anonymized creative artifacts, with no identifying information about authorship or experimental conditions. The AI is instructed to assign an overall creativity score on a 1-10 Likert scale, without receiving any additional rubric or explicit evaluation criteria beyond the prompt. To account for the stochastic nature of large language model outputs, each artifact is evaluated repeatedly over 20 independent AI runs, each implemented in a separate temporary session. This design allows us to assess both the stability of AI-based creativity judgments and the extent to which variation across runs affects reliability.

[1] We retain the terminology of Charness and Grieco (2019), though Attanasi et al. (2021), following Guilford (1975), classified these tasks as "open with constraints".

[2] In algorithm and procedure design, different algorithms or pipelines can produce the same output while varying in the number and type of computational steps, intermediate transformations, and logical structure. In patent design-around, the relevant creative act is to preserve the same user-facing function while avoiding protected claims through a genuinely different technical mechanism. In both cases, creativity lies less in changing the endpoint than in constructing a novel and well-articulated path to it.

[3] Notable exceptions include Boudreau and Lakhani (2011), who study elite programmers engaged in solving a real computational engineering problem, and Gneezy et al. (2021), who compare the performance of expert and non-expert freelancers in a closed creative task involving the generation of attractive titles for online videos.

The results of Study 1 show that AI-based creativity evaluations are highly consistent with expert human judgments. Across repeated and independent AI runs, internal consistency remains systematically high. When the AI is added to a panel of four human judges as an additional evaluator, the resulting reliability falls well within the range typically observed in CAT-based studies relying exclusively on human experts. Importantly, this result holds not only for the most favorable AI realization, but also for the least favorable one: even in the worst-performing iteration, the AI evaluator behaves similarly to a human judge and does not disrupt panel coherence. Moreover, reliability patterns vary across task domains in ways that closely mirror human evaluation. Agreement is strongest in visually constrained tasks, remains solid though more sensitive in mathematical problem-solving, and is lowest in verbal tasks, where human judgments themselves tend to be more heterogeneous. Across domains, the AI evaluator does not systematically inflate disagreement nor artificially impose consensus, but instead reproduces the domain-specific structure of human evaluative variability. Study 2 sheds light on the internal logic underlying these evaluations. When overall creativity judgments are decomposed into fluency, flexibility, originality, and elaboration, each dimension contributes positively and significantly to AI-assigned creativity scores. The relative importance of these dimensions varies systematically across drawing, mathematical, and verbal tasks, indicating that AI evaluation is both multidimensional and sensitive to task structure rather than driven by a single, domain-invariant heuristic.

Taken together, the two studies show that AI-based creativity evaluation is both reliable and substantively structured. Accounting explicitly for the stochastic nature of large language models, AI evaluations display levels of internal consistency comparable to those of human judges and integrate smoothly into human evaluation panels, even in less favorable realizations. At the same time, AI judgments are organized along theoretically grounded creativity dimensions and adapt their relative weighting to task-specific constraints, rather than being mechanically driven by prompt design or a single evaluative heuristic.

This paper contributes to the literature in four main ways.

First, we contribute to the growing literature on artificial intelligence and creativity by shifting attention from AI as a *creator* to AI as an *evaluator* of human creative work. While prior research has predominantly examined how humans perceive and evaluate AI-generated outputs (e.g., Magni et al., 2024; Arora et al., 2025; Charness and Grieco, 2026; Bellemare-Pepin et al., 2026), we study the reverse configuration, in which an AI system evaluates human creativity. By conceptualizing AI as a gatekeeping and legitimating actor in innovation processes, our analysis speaks directly to debates on algorithmic judgment, authorship, and value attribution in creative and cultural industries (Lee, 2022; Kalpokas, 2023; Oksanen et al., 2023).

Second, we advance research on algorithmic decision-making and evaluation in organizations by providing a benchmarked and conservative assessment of AI reliability relative to human judgment. While existing work on algorithmic evaluation has highlighted issues of bias, opacity, and legitimacy in domains such as hiring, credit, and performance assessment (Abrardi et al., 2022; Li et al., 2026), a growing literature on automated scoring has explored the use of machine-learning and large-language-model approaches in settings where evaluative criteria are relatively standardized, to reduce the cost, latency, and inconsistency of human scoring, while preserving alignment with domain-relevant evaluative criteria (Beaty et al., 2022; Organisciak et al., 2023; Marrone et al., 2023). We extend this literature focusing on creativity evaluation, a domain characterized by high subjectivity and ambiguity, and show that AI-based judgments can reproduce levels of internal consistency comparable to expert-based Consensual Assessment Technique benchmarks (Amabile, 1982, 1996). Once reliability is established, the data reveal three additional properties of AI-based evaluation that go beyond mere similarity to human judgment: lower evaluative variability, systematically higher scores, and task-independence of evaluative standards. Taken together, these properties motivate a hybrid model of AI-assisted evaluation, in which AI supports early-stage screening while human deliberation remains central where context and accountability matter. Importantly, by explicitly accounting for the stochastic nature of large language models, we demonstrate that these results do not hinge on a favorable realization of the model, but remain robust even in least-consistent iterations.

Third, we contribute to the creativity and innovation literature by documenting that AI-based creativity evaluations are systematically structured along established theoretical dimensions. Building on the Torrance framework (Torrance, 1974), we show that fluency, flexibility, originality, and elaboration jointly explain AI-generated creativity judgments, and that their relative importance varies across task domains. This evidence complements and extends prior work emphasizing the multidimensional and context-dependent nature of creativity in innovation processes (Amabile, 1996; Lampel et al., 2000). Our findings indicate that AI evaluators do not rely on a single heuristic or mechanically apply prompt-induced criteria, but instead implement a coherent and task-sensitive evaluative logic.

Fourth, our study contributes methodologically by combining expert-generated creative outputs, closed-task experimental designs, and repeated AI evaluations to study creativity under constraints. In doing so, we speak to a broad interdisciplinary literature – spanning educational research, psychology, psychometrics, and creativity studies – that has long examined how creative performance can be evaluated reliably, validly, and in context (Runco and Jaeger, 2012; Amabile, 1982, 1996; Bolden et al., 2020; Patston et al., 2018). This literature shows that creativity assessment is feasible only under demanding conditions: evaluative criteria must be sufficiently explicit, assessment must remain sensitive to the multidimensional nature of creativity, and rater beliefs and contextual constraints can substantially shape judgments (Bolden et al., 2020; Davies et al., 2014; Andiliou and Murphy, 2010). Related work on applied creativity in education further emphasizes that task environments do not merely constrain creativity, but also help define the conditions under which different forms of creative expression become visible and assessable across domains (Harris and Carter, 2021). More broadly, our design is consistent with a consensual and context-sensitive view of creativity assessment, according to which creativity is judged through originality together with appropriateness or usefulness within a given task and context (Amabile, 1982, 1996; Runco and Jaeger, 2012; Cseh and Jeffries, 2019). The same perspective also underlies recent work on domain-specific evaluative bias, including evidence that human evaluators may more readily associate creativity with artistic than with technical domains (Patston et al., 2018). Our paper extends this interdisciplinary conversation to a setting that has received much less attention: AI as an evaluator of human creativity in constrained, innovation-like tasks. By focusing on stylized but tightly controlled innovation-like tasks (Charness and Grieco, 2019, 2023), and on professionals operating in cultural and creative industries rather than student subjects, we provide a replicable framework for assessing AI-based evaluation in environments that approximate organizational innovation more closely than unconstrained ideation tasks. This approach responds to calls for greater external validity and task realism in experimental creativity research (Attanasi et al., 2021).

A final clarification concerns the scope and external validity of our evaluation setting. Any evaluation system – human or AI – is inevitably anchored to contemporaneous norms. Aesthetic criteria, cultural tastes, and evaluative conventions evolve over time, and works that later become celebrated may be systematically undervalued in their own era. Our study does not claim immunity from this general limitation. Rather, we adopt a more circumscribed objective: to assess whether AI-based evaluations are internally consistent, structurally coherent, and aligned with expert human judgment within a well-defined evaluative framework. In this respect, the use of closed tasks with clearly specified objectives substantially reduces the evaluative ambiguity that makes

open artistic work susceptible to radical historical reinterpretation. At the same time, the external validity of our setting is strengthened by the fact that the evaluated artifacts were produced by professionals actively operating in the cultural and creative industries, whose outputs are more likely to reflect the kinds of constraints and creative decisions that characterize real-world creative work.

The remainder of the paper is organized as follows. Section 2 describes the research design and data. Section 3 presents Study 1, which examines the reliability of AI-based creativity evaluations relative to expert human judgment. Section 4 reports Study 2, which analyzes the internal structure of AI creativity assessments across task domains. Section 5 discusses the implications of the findings for innovation and organizations, and Section 6 concludes.

2. Research design and procedures

This paper builds on the experimental data of Attanasi et al. (2026), in which participants were asked to perform one of the three creative assignments originally introduced by Charness and Grieco (2023). These assignments belong to the class of *closed creativity tasks* (Charness and Grieco, 2019), characterized by clearly specified objectives and binding constraints on admissible outputs that restrict the solution space while still allowing for substantial variation in creative performance.

Task content. Experimental participants of Attanasi et al. (2026) were asked to complete one of three creative tasks – verbal, mathematical, or drawing – using a pen-and-paper format under a strict time limit of 15 min. The three tasks were administered as follows[4]:

1. *Verbal.* Participants received the following instruction: "Choose a combination of words in the list below to write an interesting story: house, zero, forgive, curve, relevance, cow, tree, planet, ring, send". Participants were explicitly instructed that they had to use all the provided words and were free to add any other words they wished.
2. *Mathematical.* Participants were instructed as follows: "Starting from the number 27, obtain the number 6 by using at least two different numerical operations. Possible answers include: $(27 : 3)-3 = 6$, or $[(27+3) : 2-12]! = 6$". Participants were required to start from the number 27 and reach the number 6 using at least two distinct operations, while being free to use any additional numbers and operations, subject to standard arithmetic rules.
3. *Drawing.* Participants were given the instruction: "Draw a picture using any combination of shapes you like: the only constraint you have is that you must use all the following shapes". The figures supplied were: three triangles, three squares, and three circles, all of similar size. Participants were required to use all the given shapes and were allowed to include any additional shapes or elements they wished.

Each task imposed explicit constraints on admissible solutions and a clearly defined objective, making the three assignments comparable within a closed creativity framework despite their domain-specific differences.

Experimental procedure and design. The original experiment was conducted across eight pen-and-paper sessions, with 30 participants per session, for a total of 240 subjects. The sessions took place between April 2015 and May 2016 in Strasbourg (France). Half of the participants were undergraduate students in economics and management at the University of Strasbourg. The remaining half consisted of creativity experts operating in the Strasbourg region, including professionals and entrepreneurs from creative industries, postdoctoral researchers and academics specializing in the economics of creativity and innovation, as well as public officers involved in creativity-related policy design. Creativity experts participated in the experiment during two major events specifically targeted at them and organized by the University of Strasbourg: the "École d'Automne – Creativity School" in November 2015, and the "Tango & Scan – Creativity Prize" in March 2016. This setting ensured that expert participants were aware that, despite being in university classrooms, they were interacting with peers sharing similar professional backgrounds.

The experiment followed a 2×4 between-subject design, with two session types (students *vs.* experts) and four treatments. No individual participated in more than one session. In student (resp., expert) sessions, all participants were drawn from a student (resp., creativity-expert) subject pool, and this information was made public at the beginning of the experiment.

In the control (Individual) treatment, participants completed their creative task in isolation and without communication. This treatment replicated the closed-creativity-with-incentives design of Charness and Grieco (2019), based on a tournament-style incentive scheme within each task. Participants were ranked according to their creativity score within their task (*e.g.*, drawing), and experimental earnings increased with rank.[5] In the remaining three group treatments, participants still completed their creative assignments individually but were placed in groups of three and allowed to quietly discuss during the task. Each group included one participant per task category (drawing, mathematical, and verbal). The three group treatments differed according to the incentive structure governing within-group interaction.[6]

In all treatments, participants' earnings were based on their individual creativity score, as determined by peer evaluation conducted at the end of the experiment. While peer ratings served exclusively to determine experimental payoffs in the original study, all analytical results, both in the original paper and in the present study, are based on evaluations by four external judges. These judges, as will be detailed below, assessed creativity at the end of all experimental sessions under double-blind conditions with respect to session and treatment.

Human creativity evaluation. At the end of the experimental sessions, all creative outputs were evaluated by multiple independent human judges using the Consensual Assessment Technique (CAT), as introduced by Amabile (1982). In line with standard CAT procedures, four external judges with experience in creative fields independently assessed each artifact by assigning an overall creativity score on a 1–10 Likert scale, where 1 indicated "not at all creative" and 10 indicated "highly creative" (Charness and Grieco, 2019, 2023). Judges were not provided with any explicit rubric or operational definition of creativity beyond their own expert judgment, consistent with the view that creativity is best assessed through intuitive recognition by appropriate observers (Barron, 1967; Amabile, 1982).

The evaluations were conducted a few days after the final experimental session (May 2016) and were fully double blind: judges were unaware of the identity of the creators, the session type (students *vs.* experts), and the treatment condition under which each creative output had been produced. The four external judges (two male and two female) were selected to ensure diversity in background and experience with creative work. Two judges were graduate students completing master's

[4] All task instructions were administered in English and were identical to those used in Charness and Grieco (2023).

[5] Competing in cohorts of five subjects within the same task, the most creative subject received €15, the least creative €3, and the three intermediate ranks were awarded €12, €9, and €6, respectively.

[6] Specifically, there was no within-group incentive in the Group-no treatment, an incentive to equally share group earnings in the Group-coop treatment, and an incentive to compete for group earnings in the Group-comp treatment. Since these group-incentive manipulations are not the object of our paper, we do not further detail them here; we refer the reader to Section 2.4 of Attanasi et al. (2026).

theses on creativity and innovation management, one was a postdoctoral researcher with a PhD in experimental economics and expertise in experimental design, and one was a professional artist with experience as an actress and assistant director in theater, cinema, and television productions. Judges were unaware of one another's involvement in the evaluation process.[7]

Each judge received a Dropbox folder containing scanned copies of the creative assignments, organized by task domain. Specifically, all drawings were compiled into one PDF, all mathematical tasks into a second PDF, and all verbal tasks into a third PDF. Assignments were fully anonymized and labeled with sequential identification numbers within each PDF. Judges were given up to 12 h to complete the evaluations and were compensated €0.50 per assignment, for a total of 360 assignments evaluated per judge.[8] Although no explicit quality-based incentive was introduced,[9] the reliability and seriousness of the human evaluations are assessed ex post using three complementary checks discussed in Section 3.1, Table 1: inter-rater reliability, within-judge score dispersion, and the absence of mechanical evaluation patterns.

Sample selection. The present paper relies on a targeted subsample of participants consisting of creativity experts engaged in professional work in the cultural and creative industries (hereafter, cultural-and-creative-industry professionals) who completed the three aforementioned creative tasks in the original experiment. In particular, our dataset consists of 30 creative artifacts,[10] produced during the experiment by these professionals: 10 drawing tasks, 10 mathematical tasks, and 10 verbal tasks.[11]

AI-based creativity evaluation. In addition to the human evaluations from the original experiment, we assessed the same set of 30 creative artifacts using a large language model (LLM), namely ChatGPT (version 5.2). The AI-based evaluation followed a standardized prompting protocol designed to mirror the conditions under which the four human judges operated. The model was provided with the specific task instructions (drawing, mathematical, or verbal) and the corresponding outputs of cultural-and-creative-industry professionals, indicating that they were produced by human subjects, with no information about authors' identity.

As for the human judges, the AI-based evaluation was instructed to assign an overall creativity score on a 1–10 Likert scale, without receiving any additional scoring rubric or explicit evaluation criteria beyond the prompt itself. Recall that the human judges were provided with a different PDF containing the scanned artifacts for each task domain. To mirror this domain-separated evaluation, we evaluated artifacts by task domain: in each AI run, we uploaded the domain-specific set of 10 artifacts in a dedicated Temporary Chat session, so that each run comprised the 10 artifacts from a single task domain (drawing, mathematical, or verbal).

To account for the stochastic nature of LLM outputs, we repeated this procedure in 20 independent runs for each task domain. Operationally, we define *independence across runs* as follows: each run consists of a newly initiated Temporary Chat session using the same ChatGPT (version 5.2) model, with identical prompts and inputs, and with no access to prior conversation history, stored memory, or any form of cross-session state. Under this procedure, each run can be interpreted as a separate and self-contained evaluation instance, so that any variation in the scores across runs reflects the model's inherent stochasticity rather than differences in inputs or contextual carryover. Throughout the paper, the term "independent runs" refers to independence in this precise procedural sense.

Overall, this protocol involves 60 independent runs (3 task domains × 20 runs) and yields 20 AI-generated overall creativity scores for each artifact (600 artifact-level evaluations in total).

The conditions ensuring such independence are grounded in the technical properties of the evaluation environment. An OpenAI business account was created specifically for this study and was unused beforehand, with no custom instructions or personalization settings enabled. Moreover, each run was conducted in a Temporary Chat session. According to OpenAI's documentation, these sessions do not access previous conversations, do not create or rely on long-term memory, and are not used for model training (OpenAI, 2024a,b). While such chats may be temporarily retained by OpenAI for safety and abuse monitoring purposes, typically for up to 30 days (OpenAI, 2024c), this retention does not affect model behavior across sessions. Taken together, these features ensure that each run is isolated from the others, thereby supporting our operational definition of independence and minimizing potential carryover or contamination effects between successive evaluations.

Across all evaluations, we employed the same general AI assessment protocol in both Study 1 and Study 2. The two studies, however, differ in their analytical focus and in the evaluation dimensions elicited from the model. Study 1 benchmarks the reliability of AI-based creativity evaluations against human expert judgments (see Appendix A for the full prompt). Study 2 examines the internal structure of AI creativity assessments by eliciting, alongside overall creativity, multiple Torrance-inspired dimension scores (see Appendix B for the full prompt). The specific methodologies and results of each study are presented in the following two sections.

3. Study 1: AI and human evaluation of creativity

Study 1 examines whether a large language model can serve as a reliable evaluator of creativity compared to expert human judges. Creativity assessment is inherently subjective and traditionally relies

[7] Judges were also balanced in terms of age (two aged 22–25 and two aged 30–35 at the time of evaluation).

[8] In addition to the assignments evaluated in the original experiment of Attanasi et al. (2026), the same judges were also involved in the evaluation of 120 additional creative tasks produced by Vietnamese undergraduate students for a related experiment (Attanasi et al., 2019).

[9] Human judges were compensated on a per-assignment basis only; no additional payment was tied to ex-ante measures of evaluation quality. This choice is consistent with standard practice in CAT-based creativity assessment, where judges are typically paid a fixed amount and evaluation quality is examined ex post through statistical checks. By contrast, AI evaluations involved no incentive mechanism in the human sense; rather, the protocol was designed to ensure comparability by relying on standardized prompts and independent sessions.

[10] The repository is hosted on Zenodo and can be accessed at the following link: https://zenodo.org/records/18431491.

[11] In the original experimental dataset, the subsample of creativity experts professionally operating in the cultural and creative industries consisted of 35 creative artifacts: 11 drawing artifacts, 13 mathematical artifacts, and 11 verbal artifacts. For the present study, we restrict attention to 30 artifacts (10 per task domain). This restriction is motivated by a technical constraint of the AI evaluation environment. In the human CAT procedure, all artifacts belonging to a given task domain were presented together to the human judges in a single PDF file. To mirror this task-domain-level evaluation structure, our AI protocol required all artifacts from a given task domain to be uploaded and evaluated within the same AI session. However, in the ChatGPT interface/model configuration used at the time of the AI study, no more than 10 files could be uploaded within a single session. We therefore selected 10 artifacts per task domain, ensuring that the AI evaluation procedure was identical and replicable across drawing, mathematical, and verbal tasks. Specifically, we selected the first 10 outputs within each task category, ordered by their original anonymous identification codes. The remaining artifacts were excluded solely because of this technical constraint, and not on the basis of performance, content, human creativity scores, or any other evaluative criterion. As reported in footnote 15, we further verify that the selected 10/10/10 subset is not unusual in terms of human inter-rater agreement when compared with the distribution of Cronbach's alpha obtained from randomly generated task-balanced 30-observation samples drawn from the full pool of 120 observations.

on expert-based evaluations, most notably through the consensual assessment technique (CAT), which has become the gold standard in experimental creativity research (Amabile, 1982). While CAT-based evaluations have proven robust and theoretically grounded, they are also costly and time-consuming, as they require the involvement of multiple qualified judges.

Recent advances in large language models raise the question of whether AI systems can complement or augment human creativity evaluation by providing consistent and scalable assessments. However, existing studies often lack direct benchmarks against human judges or rely on loosely defined creative tasks. Study 1 addresses this gap by embedding an AI evaluator into the creativity assessment process and directly benchmarking its evaluations against those of human judges. The central question is whether human judgment can be safely replaced by AI-generated evaluation. We address this question from two complementary perspectives:

a. Does AI-generated evaluation rank cultural-and-creative-industry professionals consistently with the creativity scores assigned by human judges?
b. Does substituting one human judge with an AI evaluator preserve a level of internal consistency with the remaining evaluators that is comparable to that observed when all judges are human?

3.1. Methodology

The analysis focuses on the 30 expert-generated creative artifacts previously evaluated by four independent human judges using the CAT. As discussed in the previous section, judges assigned an overall creativity score on a 1–10 Likert-type scale, which, following standard practice in experimental studies of creativity (*e.g.*, Charness and Grieco, 2019, 2022, 2023, 2026), is used as the dependent variable in Attanasi et al. (2026).

In our setting, the judges' average score serves as the benchmark against which we compare AI-based creativity evaluations. The same artifacts were evaluated by ChatGPT (version 5.2) using the same 1–10 Likert-type scale.[12] Appendix A reports the full wording of the prompt. Specifically, in each run, the model was asked to provide one overall creativity score (1–10) for each of the 10 artifacts in the task domain. Scores were returned in a structured format (one score per artifact, in the order presented in the prompt) and were extracted for subsequent reliability and rank-correlation analyses.

Reliability is assessed using Cronbach's alpha.[13] First, we use Cronbach's alpha to assess the internal consistency of the 20 independent AI evaluations within each assignment set (drawing, mathematical, and verbal), treating the 20 AI trials as repeated raters of the same set of artifacts. We then identify, within each assignment set, the AI iteration that best represents the set of 20 evaluations and the one that is the worst representation, based on standard "alpha-if-deleted" diagnostics (*i.e.*, the iteration whose removal yields the highest *vs.* the lowest Cronbach's alpha). This step highlights the reliability and portability of our methodology: the empirical results of AI-based creativity evaluation should not hinge on which specific AI iteration is used.

Once internal consistency has been assessed and the most and least representative AI iterations have been identified for each assignment set, we compute Spearman rank correlations between the average creativity score of the four human judges and (i) the AI average score across the 20 iterations, (ii) the AI score from the most representative (Best) iteration, and (iii) the AI score from the least representative (Worst) iteration. This provides the empirical test for Research Question (a) on rank consistency between AI and human evaluations.

We then complement our analysis with the empirical test for Research Question (b) on whether replacing one human judge with an AI evaluator preserves panel reliability. Specifically, within each task domain (drawing, mathematical, and verbal), we compute Cronbach's alpha for a four-rater panel consisting of each combination of three (out of four) human judges plus the AI rating from either the most representative (Best) or the least representative (Worst) AI iteration. We compare the resulting internal consistency to the corresponding four-human baseline within the same domain. If replacing one human judge with the AI evaluator does not materially reduce Cronbach's alpha relative to the all-human benchmark, this provides evidence that AI evaluation can serve as a reliable substitute for a human judge in a CAT setting.

Table 1
Internal consistency of human judges' creativity evaluations by task domain.

Item	Drawing (α_D)	Mathematical (α_M)	Verbal (α_V)
Judge 1	0.8774	0.7196	0.6567
Judge 2	0.8571	0.5420	0.6899
Judge 3	0.8683	0.7170	0.4184
Judge 4	0.8865	0.6386	0.3820
Test scale	**0.9013**	**0.7215**	**0.6258**

Notes: The table reports Cronbach's alpha by task domain. Judge-level values correspond to *alpha if item deleted*; the last row reports the overall four-judge scale alpha.

3.2. Results

We first examine the internal consistency of AI-generated evaluations. Within each assignment set (drawing, mathematical, and verbal), the distribution of Cronbach's alpha across the 20 independent AI iterations is extremely concentrated, indicating negligible run-to-run variation in reliability. Specifically, alpha is centered around 0.9962 for the 10 drawing assignments, 0.9772 for the 10 mathematical assignments, and 0.9989 for the 10 verbal assignments, with essentially no dispersion (maximum standard deviation = 0.001 for the mathematical assignment set). Based on the "alpha-if-deleted" diagnostics described in Section 3.1, the most representative (highest-alpha) and least representative (lowest-alpha) AI iterations are, respectively, #3 and #18 for the drawing set, #4 and #20 for the mathematical set, and #17 and #20 for the verbal set. The fact that the most (Best) and least representative (Worst) iterations differ across assignment sets is consistent with the independence of our trials. Overall, internal consistency is extremely high across all three task domains.[14]

Correspondingly, we examine the internal consistency of the four human judges' creativity evaluations. In the full sample of 240 participants analyzed in Attanasi et al. (2026), Cronbach's alpha across judges is 0.7989. When we restrict attention to the 30 professionals operating in cultural and creative industries analyzed in this paper, Cronbach's alpha is 0.7759. This value is not statistically different from

[12] See Section 2 for the procedures and for the operational definition of *independence across runs*.

[13] Cronbach's alpha summarizes the internal consistency of ratings provided by multiple judges evaluating the same set of objects: higher values typically indicate stronger agreement, whereas lower values reflect more dispersed judgments.

[14] To rigorously evaluate the reliability of the AI-based evaluation framework, we implemented an exhaustive enumeration procedure that systematically explored all possible combinations of four distinct AI runs (without replacement), consistent with the fact that the human judge panel consisted of four raters. For each admissible configuration, we computed Cronbach's alpha as a standard measure of internal consistency. Across the entire distribution of resulting combinations, Cronbach's alpha values were uniformly high, consistently exceeding the conventional threshold of 0.90 and exhibiting minimal variability. This remarkable stability confirms exceptionally strong internal coherence among the AI-generated creative assessments, indicating that the evaluation process is robust to the specific selection of AI runs and that the aggregated judgments reflect a highly reliable measurement system.

the alpha computed in the full sample, nor from the corresponding alpha obtained for the remaining expert participants (*i.e.*, the other 90 experts out of 120), who do not operate professionally in cultural and creative industries.[15]

Table 1 reports Cronbach's alpha across the four human judges, separately for the 10 cultural-and-creative-industry professionals within each of the three task domains.

The pattern of internal consistency differs markedly across domains. For the drawing task, Cronbach's alpha is extremely high (above 0.9), essentially comparable to what we observe across the 20 AI evaluations. Moreover, alpha remains very high (always above 0.85) even when excluding one of the four judges from the pool. For the mathematical and verbal tasks, instead, Cronbach's alpha is substantially lower — by roughly 20 percentage points for the mathematical task and 30 percentage points for the verbal task. These values still fall within the range typically reported in CAT-based studies relying exclusively on human judges, where values above 0.5 are generally considered indicative of satisfactory reliability (Amabile, 1982, 1996; Charness and Grieco, 2019; Attanasi et al., 2021). However, especially for the verbal task, excluding one of the four judges would reduce internal consistency below the conventional threshold. Overall, internal consistency is very high for the drawing task, lower for the mathematical task, and relatively low for the verbal task. However, independently from the task, the distribution of scores assigned by each judge shows substantial within-judge variation over the 1–10 scale. Even for the judge with the lowest dispersion, scores range from 2 to 8, with a standard deviation of 1.5. This rules out mechanical scoring patterns in which the same score is repeatedly assigned to all items.

With these AI and human reliability checks in mind, we now test (a) the rank consistency between the human judges' and the AI's creativity scores. Table 2 reports three panels for the three task domains: drawing (left), mathematical (center), and verbal (right). In each panel, the table reports, for each creative worker ID, the average creativity score of the four human judges, the AI average score across the 20 iterations, the AI score from the most representative (Best) iteration, and the AI score from the least representative (Worst) iteration. At the bottom of the last three (AI) columns, it reports Spearman rank correlations with the human judges' average scores.

Table 2 also shows that AI-generated scores are generally higher than those of human judges, especially in the most technical task, the mathematical one. In this domain, AI scores are higher for all cultural-and-creative-industry professionals across the 20 iterations.

Moreover, AI-generated scores exhibit significantly lower dispersion than human evaluations across all task domains, indicating greater procedural consistency. Based on assignment-level standard deviations, paired tests reject equality of variability in favor of lower dispersion under AI evaluation (paired t-test: $p < 0.01$; Wilcoxon signed-rank test: $p < 0.01$). This holds regardless of the task, with the weakest evidence observed for the mathematical task (paired t-test: $p = 0.034$; Wilcoxon signed-rank test: $p = 0.059$). A second difference emerges in the overall level of evaluation. AI scores are significantly higher than human scores across all domains, with paired tests rejecting the null of equal means in favor of $\mu_H < \mu_{AI}$ (paired t-test $p < 0.01$; Wilcoxon $p < 0.01$).

A third pattern concerns the cross-task structure of evaluations. Human judges assign significantly lower scores to the mathematical task relative to both drawing and verbal tasks. Two-sample tests provide directional support for this pattern: mathematical assignments receive lower evaluations than drawing ones (t-test $p = 0.055$; Wilcoxon $p = 0.060$) and significantly lower evaluations than verbal ones (t-test $p < 0.01$; Wilcoxon $p < 0.01$), while no statistically significant difference emerges between drawing and verbal tasks. In contrast, AI evaluations do not exhibit any systematic variation across task domains: directional tests fail to detect significant differences between mathematical, drawing, and verbal assignments (all t-tests $p > 0.1$; Wilcoxon $p > 0.1$), suggesting that AI applies a task-invariant evaluative standard. To check the robustness of this result, we also estimate interaction models allowing the mathematical penalty to differ between human and AI evaluations. The domain-specific penalty affecting mathematical tasks in human evaluations is significantly attenuated under AI evaluation, with both the difference in the mathematical–drawing and the difference in the mathematical–verbal gap being statistically significant with, respectively, $p = 0.064$ and $p < 0.01$. Therefore, these findings suggest that domain-specific penalties on mathematical creativity are a feature of human evaluation that AI-based assessment does not replicate.

However, by construction, rank correlation is independent of the absolute score level and depends only on the ordering. In that regard, AI-generated scores exhibit very high correlations with the human judges' scores for the drawing task, with all coefficients above 0.83 and significant at the 1% level. Correlations are lower for the mathematical task, but remain statistically significant for both the most representative and the least representative AI iterations. For the verbal task, instead, the correlation is statistically significant for the AI average and for the most representative iteration, but not for the least representative one.

Importantly, the superiority of the "most representative" AI iteration is not systematic. In both the drawing and the mathematical tasks, the least representative iteration exhibits a higher rank correlation with the human benchmark than the most representative one (and the difference is sizable in the mathematical task). Moreover, when pooling all 30 cultural-and-creative-industry professionals (*i.e.*, considering the three task domains jointly), Spearman correlations with the human-judge benchmark do not differ meaningfully across the AI average score ($\rho = 0.596$), the most representative AI iteration ($\rho = 0.627$), and the least representative AI iteration ($\rho = 0.606$). All pooled correlations are significant at the 1% level. We interpret this as evidence of the reliability and portability of our methodology: the main empirical regularities do not hinge on which specific AI run is used.

More importantly, Research Question (a) shows that AI provides creativity rankings that are closely aligned with those of human judges in domains where human judges already display strong agreement under the CAT. In domains where human judges exhibit lower internal agreement, AI aligns less with the human benchmark. Research Question (b) then asks whether, in these latter domains as well, one can replace a human judge with an AI evaluator without further reducing internal consistency.

Table 3 reports Cronbach's alpha statistics to assess whether the AI evaluator can replace one human judge, using either the most representative ("Best") or the least representative ("Worst") AI iteration identified within the 20 AI trials. The table reports results for the pooled sample of 30 tasks and separately by task domain (drawing, mathematical, and verbal). For each rater i, the value reported in row i is the Cronbach's alpha computed for the remaining four raters, *i.e.*, excluding rater i ("alpha if rater deleted").[16] For completeness, Table 3 also reports the overall five-rater scale reliability (row "Test Scale"), obtained when adding the AI evaluator to the four human judges. We

[15] We conducted a stratified Monte Carlo randomization test by repeatedly drawing 30-observation samples from the full pool of 120 observations (1000 replications), requiring that each sample satisfy the composition constraint of 10 drawing, 10 mathematical, and 10 verbal tasks. The resulting distribution of Cronbach's alpha had a mean of 0.832 (SD = 0.044). The reliability of our selected sample ($\alpha = 0.7759$) did not differ statistically from the distribution of randomly generated samples (two-sided $p = 0.163$), confirming that inter-rater agreement in our subset is comparable to that in the remaining sample.

[16] The value reported when excluding the AI rater is identical across the Best and Worst columns because it is computed using the four human judges only. Furthermore, when restricting the sample to a single task domain (drawing, mathematical, or verbal), the values reported when excluding the AI rater coincide with the corresponding "Test scale" at the bottom of Table 1, which reports reliability for the four human judges only.

Table 2
Average creativity evaluations by human judges and AI.

Drawing					Mathematical					Verbal				
ID	Human	AI	B_D (#3)	W_D (#18)	ID	Human	AI	B_M (#4)	W_M (#20)	ID	Human	AI	B_V (#17)	W_V (#20)
D61	4.25	4.40	5	4	M61	4.50	6.60	6	6	V63	4.75	6.00	6	6
	(1.50)	(0.50)				(2.08)	(0.60)				(2.22)	(0.00)		
D80	7.25	7.10	7	7	M70	4.25	4.90	3	5	V70	6.25	7.00	7	7
	(1.71)	(0.31)				(2.06)	(0.79)				(1.89)	(0.00)		
D81	5.75	8.00	8	8	M76	2.00	7.55	8	4	V80	5.00	6.00	6	6
	(2.22)	(0.32)				(0.82)	(1.43)				(1.15)	(0.00)		
D83	5.75	6.20	6	6	M77	2.50	5.45	4	7	V83	4.25	5.00	5	5
	(0.96)	(0.41)				(1.91)	(1.82)				(0.96)	(0.00)		
D85	4.75	5.55	6	5	M78	6.25	8.85	9	8	V87	6.75	8.00	8	8
	(0.96)	(0.51)				(2.63)	(0.59)				(0.96)	(0.00)		
D91	7.25	8.05	8	9	M79	4.75	7.50	7	9	V89	7.00	7.85	8	7
	(2.36)	(0.76)				(2.75)	(0.69)				(2.31)	(0.59)		
D92	4.50	5.00	5	6	M83	4.50	5.80	5	6	V91	7.50	6.75	7	6
	(1.29)	(0.73)				(1.91)	(0.62)				(1.91)	(0.44)		
D94	6.25	7.35	7	7	M90	3.75	5.00	4	6	V94	7.00	4.00	4	4
	(2.06)	(0.49)				(0.50)	(0.65)				(1.15)	(0.00)		
D95	8.50	9.30	9	10	M91	7.25	8.25	9	8	V96	4.50	3.00	3	3
	(1.00)	(0.47)				(1.26)	(0.64)				(2.65)	(0.00)		
D96	2.75	5.25	6	5	M93	5.00	5.50	5	7	V97	4.25	4.80	5	5
	(0.50)	(0.64)				(0.82)	(1.15)				(0.96)	(0.41)		
ρ_s H vs. AI		**0.889**			ρ_s H vs. AI		0.541			ρ_s H vs. AI		**0.553**		
ρ_s H vs. B_D			**0.832**		ρ_s H vs. B_M			**0.555**		ρ_s H vs. B_V			**0.553**	
ρ_s H vs. W_D				**0.880**	ρ_s H vs. W_M				**0.731**	ρ_s H vs. W_V				0.438

Notes: The table reports average creativity scores for the 30 assignments, disaggregated by task domain (*drawing*, *mathematical*, and *verbal*; $N = 10$ per domain). *Human* denotes the mean score assigned by the four human judges, while AI denotes the mean score across the 20 independent AI evaluation iterations. Standard deviations of both human and AI evaluations are reported in parentheses below the corresponding mean values. Columns B_D, B_M, and B_V report the AI scores from the iteration with the highest (Best) internal consistency (Cronbach's alpha) within each domain; W_D, W_M, and W_V analogously report scores from the iteration with the lowest (Worst) internal consistency. The bottom rows report Spearman rank correlation coefficients (ρ_s) with respect to the human benchmark. Bold coefficients are statistically significant at least at the 10% level.

report this statistic as a diagnostic check, even though Research Question (b) focuses on substitution-based four-rater panels (three humans and the AI) benchmarked against the four-human baseline.[17]

At the pooled level ($N = 30$), two patterns stand out. First, focusing on the substitution design that underlies Research Question (b), the "alpha-if-deleted" entries indicate that replacing one human judge with the AI typically preserves reliability close to the four-human benchmark (obtained by excluding the AI). The four-human baseline is $\alpha = 0.7759$ (row AI). Under both the Best and Worst AI iterations, 2 out of 4 substitutions (replacing Judge 1 or Judge 3) yield reliability in line with, and in some cases weakly above, this benchmark. In the remaining cases, reliability is lower than the benchmark but remains within the range generated by omitting individual human judges, indicating that AI replacement does not generate an atypical deterioration relative to ordinary rater-to-rater variability. Second, considering the five-rater (four humans plus AI) reliability reported in the "Test Scale" row, the corresponding alpha exceeds the four-human benchmark under both the Best (0.8009) and Worst AI iteration (0.8022). This provides a useful diagnostic check that adding the AI does not undermine the internal coherence of the evaluative panel.

We next examine whether these conclusions vary across task domains. For the drawing task, reliability is exceptionally high throughout. The four-human baseline is $\alpha = 0.9013$ (row AI), and substitution-based reliability matches or exceeds this benchmark in 2 out of 4 cases under the Best iteration and in 3 out of 4 cases under the Worst iteration. Consistently, the five-rater Test Scale also exceeds 0.91 in both cases (0.9168 in Best; 0.9253 in Worst), indicating strong consensus in constrained visual-art tasks and suggesting that AI evaluation is fully compatible with the human panel in this domain.

For the mathematical task, reliability is lower than for drawing but remains strong. The four-human baseline is $\alpha = 0.7215$ (row AI). Substitution-based reliability matches or exceeds this benchmark in 1 out of 4 cases under the Best iteration and in 2 out of 4 cases under the Worst iteration. The five-rater Test Scale is again higher than the four-human benchmark and is somewhat higher in the Worst iteration (0.7521 in Best; 0.7892 in Worst), consistent with the AI aligning reasonably well with the human panel even in a more technical, rule-based domain.

For the verbal task, reliability is the lowest among the three domains, consistent with the idea that verbal creativity tends to generate more heterogeneous evaluations. The four-human baseline is $\alpha = 0.6258$ (row AI), and substitution-based reliability matches or exceeds this benchmark in 2 out of 4 cases under both the Best and Worst iterations. Thus, even in the most disagreement-prone domain, replacing one human judge with the AI does not disproportionately drive inconsistency relative to the all-human benchmark. For completeness, at the five-rater level (row "Test Scale"), reliability remains higher than the four-human baseline in both cases (Test Scale $\alpha = 0.6939$ in Best; $\alpha = 0.6675$ in Worst).

Taken together, the results addressing Research Question (b) indicate that substituting one human judge with an AI evaluator typically preserves internal consistency at levels comparable to an all-human panel. This conclusion holds in both the most and least representative AI iterations and across heterogeneous task domains. Rather than introducing additional variability, the AI evaluator appears to reproduce the shared evaluative standards that underpin consensual assessments of creativity in constrained, innovation-like tasks. Importantly, this reliability-based evidence is fully consistent with the rank-based results for Research Question (a): in domains where human judges display stronger agreement, AI aligns more closely with the human benchmark, whereas in domains with weaker human agreement, AI–human alignment is correspondingly lower.

[17] Interpreting the "Test Scale" row requires noting how Cronbach's alpha behaves when the number of raters increases. Holding constant average inter-rater agreement, α mechanically increases with the number of raters: adding an additional rater tends to raise the reliability of the composite evaluation by aggregating more independent signals. This increase is not automatic, however, because Cronbach's alpha depends jointly on the number of raters and on their average agreement. If an added rater is weakly aligned with the existing panel and lowers average agreement, the overall α may remain unchanged or even decline.

Table 3
Inter-rater reliability by rater and task domain (best and worst AI iterations).

	All tasks		Drawing (α_D)		Mathematical (α_M)		Verbal (α_V)	
Rater	Best	Worst	Best	Worst	Best	Worst	Best	Worst
Judge 1	0.7726	0.7827	0.9154	0.9206	0.6948	0.7708	0.7627	0.7226
Judge 2	0.7272	0.7328	0.8756	0.8963	0.6229	0.6935	0.6881	0.6599
Judge 3	0.7964	0.7954	0.8940	0.9102	0.7800	0.8269	0.5174	0.4853
Judge 4	0.7371	0.7285	0.9022	0.9125	0.7018	0.7151	0.5747	0.5381
AI	0.7759	0.7759	0.9013	0.9013	0.7215	0.7215	0.6258	0.6258
Test Scale	0.8009	0.8022	0.9168	0.9253	0.7521	0.7892	0.6939	0.6675

Notes: The table reports Cronbach's alpha reliability statistics for creativity evaluations based on a five-rater panel composed of four human expert judges and one AI evaluator. Columns labeled *Best* and *Worst* refer to results obtained using either the most representative (Best) or the least representative (Worst) AI iteration. Entries for individual raters report *Cronbach's α if rater deleted*, while the row "Test scale" reports the overall α including all five raters. Results are shown for the pooled sample of creative artifacts ($N = 30$) and separately by task domain (Drawing, Mathematical, and Verbal; $N = 10$ each).

4. Study 2: Creativity dimensions in AI evaluations

The second study investigates how a large language model internally structures creativity evaluations when overall creativity judgments are decomposed into multiple dimensions. While Study 1 focuses on the reliability of AI-based creativity scores relative to human expert judgments, Study 2 shifts attention to the internal logic of the AI evaluator itself. Specifically, this study examines whether overall creativity assessments generated by the AI can be explained by distinct creativity dimensions inspired by the Torrance framework, and whether the relative importance of these dimensions varies across different types of creative tasks.

Creativity is widely recognized as a multidimensional construct rather than a unitary trait. Following Torrance (1974), creativity can be assessed along four core dimensions: fluency, flexibility, originality, and elaboration. These dimensions capture complementary aspects of creative performance, ranging from idea generation and cognitive flexibility to novelty and depth of development. Study 2 leverages this multidimensional perspective to explore how an AI system operationalizes creativity when evaluating constrained creative outputs (*i.e.*, how overall creativity maps into the four dimension scores), and whether task structure shapes the relative importance of different dimensions (*i.e.*, how the contribution of fluency, flexibility, originality, and elaboration to overall creativity varies across drawing, mathematical, and verbal tasks).

4.1. Methodology

Study 2 relies on the same set of 30 expert-generated creative assignments described in Section 2. As in Study 1, all evaluations are conducted exclusively by ChatGPT (version 5.2), following the same standardized AI protocol and independent-trial procedure described in Section 2. Human judges are not involved in Study 2; instead, the analysis focuses on the internal consistency and explanatory structure of AI-generated creativity judgments.

Unlike Study 1, in each trial the model was asked to provide not only an overall creativity score but also separate evaluations along four creativity dimensions inspired by Torrance (1974): fluency, flexibility, originality, and elaboration. Overall creativity was evaluated on a 1–10 Likert-type scale, consistent with Study 1, while each Torrance dimension was evaluated on a 1–5 Likert-type scale for each artifact presented in the prompt. Results were returned in a single table. The full wording of the prompt, including scale definitions and output constraints, is reported in Appendix B.

For each dimension, the AI was provided with a brief conceptual definition designed to guide the evaluation without imposing rigid scoring rules. The definitions we provided capture the core Torrance dimensions: *fluency* as the number of relevant ideas, *flexibility* as the diversity of idea categories or shifts in approach, *originality* as the relative unusualness of the response compared to typical outputs, and *elaboration* as the amount of detail and development. In this regard, while the Torrance Tests of Creative Thinking (TTCT) scoring is typically norm- and count-based (especially for originality),[18] we operationalize these constructs via Likert-type ratings to elicit comparable multidimensional judgments from the LLM.

As in Study 1, to account for the stochastic nature of LLM outputs, each creative artifact was evaluated in 20 independent trials. Each trial was conducted in a new Temporary Chat session using the same model version and an identical prompt, with no access to prior conversation history or stored memory. This procedure mirrors Study 1 and yields 600 AI-generated evaluations (10 assignments × 20 repetitions × 3 tasks). For each evaluation, the AI returned a complete vector of scores, including an overall creativity rating (comparable to Study 1) and four Torrance-inspired dimension ratings.

The empirical analysis examines how overall creativity relates to the four dimension scores using regression methods, and assesses whether these relationships differ across task domains. Formally, we estimate:

$$Creativity_{it} = \alpha + \beta_1 Fluency_{it} + \beta_2 Flexibility_{it} + \beta_3 Originality_{it} + \beta_4 Elaboration_{it} + \gamma_t + \delta_i + \epsilon_{it}. \quad (1)$$

where $Creativity_{it}$ denotes the overall creativity score assigned by the AI to creative artifact i in task domain t; δ_i captures assignment fixed effects; and ϵ_{it} is an idiosyncratic error term. We estimate variants of the model including either task-domain fixed effects or assignment fixed effects. The model is estimated on pooled AI-generated evaluations, treating repeated ratings as independent observations.

In addition, we estimate the model separately for drawing, mathematical, and verbal tasks to explore task-specific (t) patterns in the structure of AI creativity judgments.

4.2. Results

Before turning to the regression analysis, we report two brief diagnostic checks. First, we verify that AI evaluations remain internally consistent across the 20 independent trials, as in Study 1. Second, we verify that eliciting dimension scores in the prompt does not materially distort the overall creativity scores relative to Study 1.

Regarding the first check, Cronbach's alpha is again extremely concentrated across the 20 trials within each task domain, indicating negligible run-to-run variation in reliability. Specifically, alpha is centered around 0.9943 for the 10 drawing assignments, 0.9836 for the 10

[18] In the TTCT, these dimensions are typically operationalized using standardized scoring procedures. In particular, *fluency* and *flexibility* are commonly scored as counts of, respectively, the number of relevant responses and the number of distinct response categories; *elaboration* is scored based on the amount of added detail or development beyond a basic response; and *originality* is usually scored in a norm-referenced way, assigning higher scores to statistically infrequent responses relative to TTCT scoring norms.

Table 4
Differences in AI creativity scores between Study 1 and Study 2.

Pair	Δ Mean	*p* (Mean)	*F* (Var)	*p* (Var)	Interpretation
AI(1) *vs.* AI(2)	+0.33	0.134	0.747	0.438	Weak evidence
$AI(1)_D$ *vs* $AI(2)_D$	−0.10	0.591	1.223	0.769	No difference
$AI(1)_M$ *vs* $AI(2)_M$	+0.70	0.257	0.773	0.708	No difference
$AI(1)_V$ *vs* $AI(2)_V$	+0.40	0.104	0.560	0.400	Marginal
B(1) *vs.* B(2)	+0.37	0.177	0.735	0.411	Weak evidence
$B(1)_D$ *vs.* $B(2)_D$	0.00	1.000	1.000	1.000	Identical
$B(1)_M$ *vs.* $B(2)_M$	+0.70	0.354	1.308	0.695	No difference
$B(1)_V$ *vs.* $B(2)_V$	+0.40	0.223	0.426	0.219	No difference
W(1) *vs.* W(2)	0.00	1.000	1.000	1.000	Identical
$W(1)_D$ *vs.* $W(2)_D$	+0.20	0.509	2.214	0.252	No difference
$W(1)_M$ *vs.* $W(2)_M$	−0.50	0.322	1.207	0.784	No difference
$W(1)_V$ *vs.* $W(2)_V$	+0.30	**0.081**	0.620	0.488	Marginal

Notes: The table reports paired tests comparing AI-generated overall creativity scores between Study 1 (overall creativity only) and Study 2 (overall creativity elicited jointly with Torrance-inspired dimensions). Δ Mean reports the average within-assignment difference between the two studies. *p* (Mean) refers to two-sided paired *t*-tests for equality of means, while *F* (Var) and *p* (Var) report F-tests for equality of variances. Results are shown for the pooled sample of creative artifacts and separately by task domain (Drawing, Mathematical, and Verbal). Labels *Best* (B) and *Worst* (W) denote, respectively, the AI iterations associated with the highest and lowest internal consistency within each study. These iterations are identified either on the pooled sample across all tasks or within each task domain separately (Drawing only, Mathematical only, Verbal only). Marginal significance is defined as $p < 0.10$.

mathematical assignments, and 0.9983 for the 10 verbal assignments, with essentially no dispersion (maximum standard deviation = 0.001 for the mathematical assignment set). As in Study 1, the Best and Worst iterations differ across task domains, consistent with the independence of our trials.[19]

The second check proceeds in two steps. First, we compare the average, most representative (Best), and least representative (Worst) AI-generated overall creativity scores in Study 1 *vs.* Study 2. Table 4 reports mean differences with associated *t*-tests, as well as *F*-tests for differences in variances between Study 1 and Study 2 creativity scores. When using the average creativity score, we find no statistically significant differences, either at the pooled 30-assignment level or by task domain. The same holds when comparing the Best iterations across the two studies. Consistent with our conservative approach, even when comparing the Worst iterations we find at most a marginal mean difference for the verbal task, with no differences for the other two domains and no difference at the pooled 30-assignment level. Overall, these results suggest that the two studies yield very similar distributions of overall creativity scores in levels.

Second, and more importantly, we test whether the ranking of the 30 creative assignments is invariant across studies. Table 5 reports Spearman rank correlations between Study 1 and Study 2 overall creativity scores, computed separately for the average, Best, and Worst scores. The results are reassuring: correlations are uniformly high (always above 0.75) and statistically significant at the 1% level, regardless of whether we compare average-to-average, Best-to-Best, or Worst-to-Worst scores. Strikingly, correlations remain high and significant even when the selection principle differs across studies (e.g., Best in Study 1 *vs.* Worst in Study 2), indicating that the relative ordering of assignments is highly stable.

Taken together, these diagnostics indicate that the overall creativity scores elicited under the multidimensional prompt are in line with those obtained in Study 1. With these two positive checks in mind, we now turn to our dimension-based creativity model of Study 2.

[19] Based on the "alpha-if-deleted" diagnostics described in Section 3.1, the most representative (highest-alpha) and least representative (lowest-alpha) AI iterations are, respectively, #11 and #8 for the drawing set, #13 and #2 for the mathematical set, and #17 and #20 for the verbal set.

Table 5
Spearman rank correlations between AI creativity scores in Study 1 and Study 2.

	AI(1)	AI(2)	B(1)	B(2)	W(1)	W(2)
AI(1)	1.0000					
AI(2)	0.7867***	1.0000				
B(1)	0.9381***	0.7342***	1.0000			
B(2)	0.7672***	0.9394***	0.7679***	1.0000		
W(1)	0.7758***	0.8311***	0.6795***	0.8018***	1.0000	
W(2)	0.8807***	0.7004***	0.8181***	0.6759***	0.7652***	1.0000

Notes: The table reports Spearman rank correlations between AI-generated overall creativity scores obtained in Study 1 and Study 2. AI denotes the average creativity score across the 20 independent AI iterations. Best (B) and Worst (W) denote, respectively, the AI iterations associated with the highest and lowest internal consistency (Cronbach's alpha) within each study. These iterations are identified on the pooled sample across all task domains. Correlations are computed on the pooled sample of assignments ($N = 30$). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 6
Creativity dimensions and overall creativity: Pooled AI evaluations.

	Overall creativity		
	(1)	(2)	(3)
Fluency	0.535***	0.552***	0.378***
	(0.037)	(0.036)	(0.050)
Flexibility	0.386***	0.376***	0.361***
	(0.051)	(0.051)	(0.051)
Originality	0.493***	0.487***	0.634***
	(0.038)	(0.037)	(0.039)
Elaboration	0.402***	0.387***	0.441***
	(0.038)	(0.038)	(0.045)
Task FE	N	Y	N
Assignment FE	N	N	Y
Observations	600	600	600
R-squared	0.944	0.947	0.957

Notes: The table reports OLS estimates of the relationship between AI-assigned overall creativity scores and the four Torrance-inspired creativity dimensions. The dependent variable is overall creativity evaluated by ChatGPT on a 1–10 Likert-type scale. Fluency, flexibility, originality, and elaboration are evaluated on 1–5 Likert-type scales. Column (1) reports pooled estimates without fixed effects. Column (2) includes task fixed effects. Column (3) includes assignment fixed effects. Standard errors are heteroskedasticity robust and reported in parentheses. All regressions are estimated on pooled AI-generated evaluations, treating repeated ratings as independent observations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 6 reports the baseline regression results pooling all AI-generated evaluations across tasks. Overall creativity evaluations are positively and strongly associated with all four Torrance-inspired dimensions (fluency, flexibility, originality, and elaboration) across specifications. The estimated coefficients are large in magnitude and highly statistically significant, indicating that the AI systematically integrates multiple dimensions of creative performance when forming overall creativity judgments.

Originality and fluency exhibit particularly strong associations with overall creativity in the pooled specifications, suggesting that both novelty-related features and the capacity to generate multiple relevant ideas play a central role in the AI's evaluative process. Flexibility and elaboration also contribute significantly, highlighting the importance of cognitive shifts across categories and depth of development in shaping overall assessments. The inclusion of task fixed effects and assignment fixed effects does not qualitatively alter these results, and the models display very high explanatory power, with R-squared values exceeding 0.94 across specifications. Taken together, these findings indicate that AI-based creativity evaluations are highly structured and closely aligned with a multidimensional conception of creativity.

Table 7 presents regression results estimated separately for drawing, mathematical, and verbal tasks, revealing systematic task-specific differences in the relative weighting of creativity dimensions. For drawing tasks, all four dimensions are positively associated with overall creativity, with elaboration and originality displaying particularly large

Table 7
Creativity dimensions and overall creativity by task domain.

	Overall creativity					
	Drawing task		Mathematical task		Verbal task	
	(1)	(2)	(3)	(4)	(5)	(6)
Fluency	0.402***	0.285***	0.535***	0.378***	0.688***	0.516***
	(0.077)	(0.094)	(0.047)	(0.076)	(0.074)	(0.084)
Flexibility	0.244***	0.277***	0.489***	0.420***	0.341***	0.306***
	(0.081)	(0.076)	(0.091)	(0.100)	(0.069)	(0.077)
Originality	0.464***	0.570***	0.455***	0.667***	0.586***	0.578***
	(0.052)	(0.073)	(0.070)	(0.067)	(0.054)	(0.069)
Elaboration	0.559***	0.493***	0.362***	0.465***	0.245***	0.314***
	(0.061)	(0.083)	(0.070)	(0.075)	(0.058)	(0.066)
Assignment FE	N	Y	N	Y	N	Y
Observations	200	200	200	200	200	200
R-squared	0.932	0.939	0.927	0.942	0.971	0.975

Notes: The table reports OLS estimates of the relationship between AI-assigned overall creativity scores and the four Torrance-inspired creativity dimensions, estimated separately by task domain. The dependent variable is overall creativity evaluated by ChatGPT on a 1–10 Likert-type scale. Fluency, flexibility, originality, and elaboration are evaluated on 1–5 Likert-type scales. Columns (1)–(2) refer to drawing tasks, columns (3)–(4) to mathematical tasks, and columns (5)–(6) to verbal tasks. For each task domain, specifications without assignment fixed effects are reported in odd-numbered columns, while even-numbered columns include assignment fixed effects. Standard errors are heteroskedasticity robust and reported in parentheses. All regressions are estimated on pooled AI-generated evaluations, treating repeated ratings as independent observations. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

coefficients. This pattern suggests that, in visually constrained tasks, the AI rewards both the refinement of visual composition and the distinctiveness of the solution.

For mathematical tasks, originality and fluency emerge as especially salient predictors of overall creativity, alongside a strong contribution of flexibility and elaboration. This result is consistent with an interpretation of mathematical creativity as combining novel transformations with the ability to generate multiple valid solution paths under strict constraints. The balanced importance of all four dimensions indicates that the AI evaluates mathematical creativity as a multifaceted problem-solving process rather than a purely technical exercise.

For verbal tasks, fluency and originality play a dominant role, while flexibility and elaboration also remain positive and statistically significant. Compared to the other task domains, verbal assignments exhibit the highest explanatory power, suggesting a particularly close alignment between the Torrance dimensions and the AI's overall creativity assessments in language-based tasks. This pattern reflects the centrality of idea generation and narrative novelty in verbal creativity, even within a closed-task framework.

A final implication of these task-specific results is that AI-based creativity evaluation may still privilege works that perform reasonably well across multiple dimensions over works that are radically strong on only one. This point is especially relevant for verbal creativity. Although originality is among the strongest predictors of AI-assigned overall creativity in verbal tasks, fluency also carries substantial weight. As a result, a deliberately sparse or formally simple output may be undervalued if it scores very highly on originality but relatively low on fluency or elaboration. In this sense, the multidimensional structure of AI evaluation is not equivalent to a mechanical preference for complexity, but it may still fail to fully recognize forms of creativity whose distinctiveness lies precisely in radical simplicity.

Overall, the results of this study indicate that AI-based creativity evaluations exhibit a coherent and fully multidimensional structure that closely mirrors established theoretical frameworks of creativity. Rather than privileging a single criterion, the AI integrates fluency, flexibility, originality, and elaboration in a systematic manner, with task structure shaping their relative importance. These findings suggest that, in constrained creative environments, AI evaluators operationalize creativity in a way that is both theoretically grounded and sensitive to domain-specific features of creative production.

5. Discussion

This paper examines whether a large language model can serve as a reliable and informative evaluator of human creativity in constrained, innovation-like environments. By embedding ChatGPT into a standardized creativity assessment framework and benchmarking its evaluations against expert human judges, we provide evidence on both the reliability and the internal structure of AI-based creativity judgments.

Across two complementary studies, our results show that AI evaluations of creative work are (i) internally consistent and comparable to expert human judgments, even when accounting for variability across independent AI evaluations, and (ii) systematically structured along multiple creativity dimensions whose relative importance varies across task domains.

5.1. Reliability and legitimacy of AI-based creativity evaluation

The findings of Study 1 provide evidence that AI-based creativity evaluations can reach levels of reliability comparable to those obtained through expert human judgment using the Consensual Assessment Technique. Across repeated and independent evaluation runs, the inclusion of an AI evaluator alongside human judges does not undermine inter-rater consistency. Importantly, this result holds not only on average, but also under deliberately conservative conditions, when comparing the most and the least internally consistent AI evaluations observed across runs. Even in the least favorable case, AI-based evaluations remain within accepted benchmarks for expert-based creativity assessment and perform on par with typical human raters.

Beyond reliability, the data reveal three additional features of AI-based evaluation that are difficult to achieve with human-only panels. First, AI-generated scores display significantly lower dispersion than human scores across all task domains, indicating greater procedural consistency and reduced rater drift. Second, AI scores are significantly higher than human scores across all domains and all independent iterations, consistent with the idea that AI may be less anchored to the conservative professional norms that have historically led to undervaluation of unconventional creative work — though we acknowledge this could also reflect differences in scale usage or evaluative generosity, and flag it as an open question for future research. Third, AI evaluations are essentially task-independent in their internal consistency, whereas human judges display lower agreement and lower average scores for the mathematical task relative to drawing and verbal tasks, consistent with a well-documented "art bias" in human evaluation (Patston et al., 2018). Overall, these findings motivate a hybrid model of AI-assisted evaluation: AI supports early-stage screening and provides an auditable complement to expert judgment, while human deliberation remains indispensable where context, contested novelty, or institutional accountability are central.

From an innovation and organization theory perspective, these results speak less to the technical accuracy of the AI and more to its potential *legitimacy* as an evaluative actor. In organizational settings, evaluative systems acquire authority not only because they are accurate, but because they are perceived as consistent, stable, and aligned with shared standards. The fact that AI evaluations align closely with the consensual structure of expert judgment suggests that large language models can internalize and reproduce the implicit evaluative norms that underpin expert-based creativity assessment. Crucially, this alignment is robust across independent AI evaluations, indicating that legitimacy does not hinge on selecting a particular model realization. This is particularly relevant in innovation contexts where evaluation often functions as a bottleneck. Screening creative ideas typically requires the coordination of multiple evaluators, involves substantial time costs, and is subject to disagreement and noise. An AI-based evaluator that behaves similarly to expert judges – without systematically inflating disagreement or introducing idiosyncratic variation – may therefore act

as a credible component of evaluative routines, especially in early-stage assessment.

Importantly, the results also highlight that reliability is not uniform across domains. Lower agreement in verbal tasks mirrors well-documented patterns in human evaluation, suggesting that AI does not artificially homogenize judgments in domains where creativity is inherently more interpretative. In this sense, AI-based evaluation appears to reflect, rather than override, domain-specific uncertainty. This further reinforces the view that AI operates within existing evaluative structures, rather than imposing an exogenous or domain-invariant standard of creativity.

5.2. The structure of AI creativity judgments

The results of Study 2 indicate that AI-based creativity evaluations are not ad hoc or idiosyncratic, but exhibit a clear internal structure that closely mirrors established theoretical conceptions of creativity. Across tasks, overall creativity judgments are systematically related to fluency, flexibility, originality, and elaboration, suggesting that the AI does not rely on a single heuristic or superficial cue, but instead integrates multiple dimensions of creative performance. This finding is important not because it shows that AI can "recognize creativity" in an abstract sense, but because it suggests that AI evaluation operates by formalizing and stabilizing evaluative criteria that are often implicit, contested, or unevenly applied in organizational settings. In this respect, AI-based evaluation does not introduce a novel definition of creativity; rather, it renders operational a multidimensional framework, closely aligned with the Torrance tradition, that has long informed expert judgment without being explicitly codified.

At the same time, the structure of AI judgments is not invariant across contexts. The relative weight assigned to different creativity dimensions varies systematically across task domains. This task heterogeneity does not reflect inconsistency or noise; instead, it indicates that the AI applies a common conceptual framework of creativity while calibrating its evaluative emphasis to the structure of the problem at hand. In other words, the AI appears to combine standardization at the level of dimensions with flexibility at the level of their relative importance.

From an innovation and organization theory perspective, this pattern is particularly informative. In organizational innovation processes, creativity is rarely evaluated independently of task characteristics. Different forms of innovation, such as product design, process optimization, or symbolic and narrative production, tend to privilege different aspects of creative performance. The evidence presented here suggests that AI-based evaluators reproduce this contextual sensitivity in a systematic and scalable manner.

More broadly, these findings highlight the role of AI as a technology of evaluation rather than creation. By embedding a structured and multidimensional evaluative logic into a scalable system, AI-based evaluation may contribute to the standardization of creativity assessment across heterogeneous domains. While such standardization can enhance comparability and transparency, it also implies that certain forms of creativity may become more legible, and therefore more likely to be selected, than others. In this sense, AI-based evaluation has the potential not only to assess creativity, but also to shape innovation trajectories by influencing which creative attributes are systematically rewarded.

5.3. Implications for organizations, innovation, and creative work

Taken together, the results of Study 1 and Study 2 suggest that AI-based creativity evaluation should be understood as an organizational technology of judgment rather than merely as a measurement tool. The evidence shows that AI can simultaneously achieve reliability comparable to expert human evaluators and implement a structured, multidimensional conception of creativity that is sensitive to task characteristics. This combination is particularly consequential in organizational and innovation settings, where evaluation plays a central role in shaping selection, coordination, and resource allocation.

From an organizational perspective, evaluation is not a neutral act: it defines standards of quality, establishes legitimacy, and structures decision-making under uncertainty. The fact that AI evaluations align closely with the consensual structure of expert judgment (Study 1) suggests that large language models can function as legitimate evaluative actors within existing institutional frameworks. Rather than imposing an external or idiosyncratic notion of creativity, the AI appears to internalize and reproduce the shared evaluative norms that underpin expert-based assessment. This makes AI-based evaluation potentially compatible with organizational routines that rely on expert consensus, such as peer review, commissioning, screening, and internal selection committees. At the same time, existing evidence on algorithmic aversion suggests that experts often resist delegating judgment to algorithmic systems, even when those systems perform comparably to human raters (Burton et al., 2020). This resistance may be particularly pronounced in creativity assessment, where professional identity is closely tied to evaluative authority. Whether reporting dimension-level scores alongside overall ratings – which makes the basis of the AI judgment explicit and auditable – can mitigate this resistance is an open and practically important question for future research.

At the same time, Study 2 shows that AI-based evaluation is not monolithic. While the underlying structure of creativity judgments reflects established dimensions (fluency, flexibility, originality, elaboration), the relative importance of these dimensions varies systematically across tasks. This task heterogeneity is particularly relevant for innovation theory, where creativity is widely understood as context-dependent and shaped by problem structure. Different innovation activities, such as product design, process optimization, or symbolic and cultural production, entail different evaluative priorities. The AI's ability to adjust the weighting of creativity dimensions across tasks suggests that algorithmic evaluation can replicate, at scale, forms of contextual sensitivity that are typically associated with human expertise.

These features have important implications for innovation processes. In many organizations, early-stage innovation involves screening large numbers of heterogeneous ideas under time and resource constraints. AI-based evaluation may therefore function as a scalable screening device that preserves multidimensional judgment while reducing variability and coordination costs. However, because evaluation criteria influence creative behavior, the deployment of AI evaluators is likely to generate feedback effects. As cultural-and-creative-industry professionals and teams anticipate algorithmic assessment, they may adapt their output to align with the dimensions that are systematically rewarded. In this sense, AI-based evaluation does not merely select among ideas; it may also shape the direction of creative effort and, over time, the trajectory of innovation.

The findings also speak to broader questions about creative labor and professional identity. In cultural and creative industries, value creation depends not only on production but also on legitimation through evaluative infrastructures such as juries, editors, curators, and commissioning bodies. When parts of this evaluative authority are delegated to AI systems, standards of creativity become increasingly embedded in technical artifacts rather than residing exclusively within professional communities. This shift may affect how creative expertise is recognized, how authorship is validated, and how legitimacy is conferred, especially in settings where algorithmic evaluation is used as a gatekeeping device. Creators and artists may hold particularly ambivalent attitudes toward AI-based assessment: while AI evaluation offers greater transparency and consistency relative to opaque human panels, it may also be perceived as delegitimizing by those who value recognition by a human expert. Whether dimension-level explanations increase perceived fairness among creators – and whether this varies

across domains and institutional contexts – is a question that future experimental work could address directly.

Finally, the results raise governance considerations concerning transparency and accountability. Because AI-based creativity judgments are structured and multidimensional, relying solely on aggregate scores risks obscuring the underlying evaluative logic. Reporting dimension-level assessments alongside overall ratings could enhance interpretability and support more informed organizational decision-making. Moreover, the possibility of repeated, independent AI evaluations enables systematic auditing of evaluator stability and sensitivity, offering organizations tools to assess when AI-based judgment is appropriate and when human deliberation remains indispensable. Hybrid evaluative systems, combining algorithmic screening with human oversight, may therefore represent a pragmatic approach to managing creativity and innovation in increasingly data-intensive organizational environments. More broadly, the question of whether AI-based evaluation should be adopted involves not only measurement properties and stakeholder acceptance, but also questions of fairness, power, and the distribution of evaluative authority — questions that lie at the intersection of technology governance, labor relations, and cultural policy, and that deserve sustained attention from researchers, practitioners, and policymakers alike.

An important dimension that future research should address is the acceptance of AI-based evaluation by real-world stakeholders. The practical relevance of AI-assisted evaluation depends not only on its measurement properties, but also on whether it is accepted as legitimate by the actors who would be affected by its deployment. Human evaluators may resist delegating judgment to algorithmic systems, especially in domains where professional identity is closely tied to evaluative authority. Creators may welcome AI evaluation as more transparent and consistent, or may perceive it as delegitimizing if they value recognition by human experts. Managers and institutional decision-makers may adopt AI-assisted evaluation only if they perceive it as consistent with their accountability obligations. Mapping these heterogeneities – across stakeholder groups and institutional contexts – is an important direction for future empirical work, including incentivized choice experiments and survey-based studies of evaluator and creator attitudes toward algorithmic assessment.

6. Conclusions

This paper investigates whether a large language model can act as a reliable and informative evaluator of human creativity in constrained, innovation-like tasks. Rather than focusing on AI as a creative agent, we study AI as an evaluative actor and assess the reliability, internal structure, and task sensitivity of its creativity judgments.

The empirical analysis yields three main results. First, AI-based creativity evaluations exhibit levels of internal consistency comparable to those obtained from expert human judges using the Consensual Assessment Technique. Across repeated and independent evaluation runs, adding an AI evaluator to a human judging panel does not reduce inter-rater reliability and, in some cases, performs on par with the most consistent human evaluators. Importantly, this conclusion holds even under conservative conditions, when considering the least favorable AI realizations. This indicates that AI-based evaluations are not driven by stochastic noise, but align closely with the shared evaluative standards underlying expert judgment.

Second, AI creativity judgments display a clear and coherent internal structure. When overall creativity scores are decomposed into fluency, flexibility, originality, and elaboration, all four dimensions contribute positively and significantly to overall evaluations. This finding suggests that the AI does not rely on a single heuristic or superficial indicator of creativity, but instead integrates multiple, theoretically grounded dimensions that have long been central to creativity research.

Third, the relative importance of these dimensions varies systematically across task domains. Drawing, mathematical, and verbal tasks exhibit distinct weighting patterns, reflecting differences in task structure and constraints. This task heterogeneity mirrors patterns observed in human creativity evaluation and indicates that AI judgments are context-sensitive rather than mechanically uniform across domains.

Taken together, these results have direct implications for organizations engaged in innovation and creative work. Creativity evaluation is a central bottleneck in many innovation processes, particularly in early-stage idea screening, where large numbers of heterogeneous proposals must be assessed under time and resource constraints. The evidence presented here suggests that AI-based evaluators can function as scalable screening devices that preserve multidimensional judgment while maintaining reliability comparable to expert human assessment. When used as a complement to human judgment, AI evaluation can reduce coordination costs and variability without imposing a rigid or domain-invariant notion of creativity.

At the same time, evaluation systems shape creative behavior. As cultural-and-creative-industry professionals anticipate algorithmic assessment, they may adapt their output to align with the dimensions that are systematically rewarded. This implies that AI-based evaluation does not merely select among ideas, but may also influence the direction of creative effort and, over time, the trajectory of innovation within organizations. The design and governance of AI evaluation systems therefore become managerial choices with strategic consequences.

The findings are also relevant for policy contexts in which creativity evaluation plays a gatekeeping role, such as cultural funding, innovation grants, startup competitions, and public procurement. In these settings, consistency, transparency, and procedural fairness are key concerns. The evidence that AI-based evaluations are reliable and multidimensional suggests that such systems could support standardized preliminary screening while reducing evaluator workload. At the same time, reliance on aggregate creativity scores alone risks obscuring the underlying evaluative logic. Reporting dimension-level assessments alongside overall ratings could enhance transparency and accountability, allowing applicants and policymakers to better understand selection decisions. Moreover, repeated and independent AI evaluations enable auditing of evaluator stability, offering new tools for governance and oversight.

Several limitations qualify the scope of our conclusions and point to directions for future research. First, the creative tasks analyzed here are deliberately stylized. While closed tasks capture important features of innovation under constraints, they abstract from interaction, iteration, and organizational dynamics that often shape creative production and evaluation in practice. Field studies in organizational settings – examining whether AI-based evaluation of constrained creative tasks predicts subsequent innovation outcomes – would help assess the external validity of AI-based creativity evaluation. Second, although the Torrance-inspired dimensions provide a well-established framework, our Likert-type operationalization differs from norm- and count-based scoring procedures used in standardized creativity tests, particularly for originality. Future work could explore alternative scoring protocols and assess the robustness of the observed evaluative structure. Third, AI evaluation is inherently model- and prompt-dependent. Replications across different models, prompts, and evaluation settings would be valuable to map the stability and boundaries of AI-based creativity judgments.

Fourth, a fundamental boundary condition on AI-based evaluation is its anchoring to contemporaneous norms. AI systems learn from the corpus of human-generated content available at training time and cannot engage in the kind of long-run revision of evaluative standards that human communities of practice are capable of. This implies that AI evaluators may reproduce the conformist biases embedded in the training distribution, without the capacity for the generational reappraisal that has historically rehabilitated undervalued creative work. A related concern is that AI systems which efficiently aggregate existing human judgments may generate mean-attracted evaluations that systematically disadvantage genuinely unconventional work — a limitation that our

Study 2 results only partially address. While the strong weight placed on originality across all task domains suggests that the AI does not simply reward conventional outputs, a work scoring very low on fluency or elaboration while being exceptional on originality may still receive a lower overall score than it deserves. Future research might explore evaluation protocols that explicitly flag such outlier profiles – artifacts with unusually high originality but low scores on other dimensions – as candidates for deeper human deliberation, precisely because they may embody the kind of unconventional creativity that aggregative systems tend to undervalue.

Fifth, a further limitation concerns the scalability of the evaluation protocol. The current design was constrained by the maximum number of files that could be uploaded within a single AI session at the time of our study, which restricted our analysis to 10 artifacts per task domain. This is a technical boundary of the specific model version and interface used rather than an intrinsic feature of AI-based evaluation. As large language model capabilities continue to evolve, with newer systems increasingly supporting larger context windows and file upload capacities, this constraint is likely to become less binding, enabling future replications to evaluate larger, and if needed unbalanced, artifact sets.

Finally, the fact that AI-generated scores in our data are systematically higher than human scores across all task domains and iterations should be interpreted cautiously. Although this pattern may suggest that AI is less prone to the kind of conservative undervaluation that has historically characterized some expert panels, and may therefore point to a genuinely more inclusive evaluative stance, it could also reflect differences in scale use or evaluative generosity. This remains an open question for future research.

CRediT authorship contribution statement

Valerio Fedele Addis: Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Giuseppe Attanasi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Giovanni Di Bartolomeo:** Writing – review & editing, Investigation, Conceptualization. **Michele Mariella:** Writing – original draft, Software, Methodology, Formal analysis, Data curation. **Valentina Peruzzi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Project administration, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. AI prompt for creativity evaluation (Study 1)

This appendix reports the prompt used to elicit overall creativity evaluations from ChatGPT in Study 1. The same prompt was used across all independent evaluation trials. No additional instructions, examples, or scoring rubrics were provided beyond the text reported below.

Prompt text:

> On a Likert scale from 1 to 10, how creative was each of the above assignments. Create a table and answer by indicating only the table ID and the scores.

The model was provided with the original task instructions and the corresponding human-generated solution before receiving the evaluation prompt. In our implementation, the "table ID" corresponds to the assignment identifier used to match model outputs to artifacts in the dataset. Each evaluation was conducted in a new Temporary Chat session to ensure independence across repetitions.

Appendix B. AI prompt for creativity dimensions (Study 2)

This appendix reports the prompt used to elicit multidimensional creativity evaluations from ChatGPT in Study 2. The prompt was designed to decompose overall creativity into four dimensions inspired by the Torrance framework: fluency, flexibility, originality, and elaboration. The same prompt was used across all independent evaluation trials.

Prompt text:

> Following Torrance (1974), creativity can be assessed along four core dimensions: fluency, flexibility, originality, and elaboration.
>
> Please evaluate each of the assignments above using the Likert-type scales and definitions below.
>
> Overall creativity (1–10): 1 = not creative at all; 10 = extremely creative.
>
> Fluency (1–5): number of distinct relevant ideas (1 = very low; 5 = very high).
>
> Flexibility (1–5): number of shifts in thinking or distinct response categories (1 = very low; 5 = very high).
>
> Originality (1–5): how unusual/rare the response is relative to typical responses (1 = not unusual; 5 = highly unusual).
>
> Elaboration (1–5): amount of detail and development of the idea (1 = very low; 5 = very high).
>
> Output format (no extra text): Return only a table with one row per assignment and the following columns:
>
> Assignment_ID in ascending order | Overall_1_10 | Fluency_1_5 | Flexibility_1_5 | Originality_1_5 | Elaboration_1_5

The model received the original task instructions and the corresponding human-generated solution prior to evaluation. In our implementation, the "Assignment ID" corresponds to the assignment identifier used to match model outputs to artifacts in the dataset. Each evaluation was conducted in a new Temporary Chat session to ensure independence across repetitions.

Data availability

Data will be made available on request.

References

Abrardi, L., Cambini, C., Rondi, L., 2022. Artificial intelligence, firms and consumer behavior: A survey. J. Econ. Surv. 36 (4), 969–991.

Amabile, T.M., 1982. Social psychology of creativity: A consensual assessment technique. J. Pers. Soc. Psychol. 43 (5), 997–1013.

Amabile, T.M., 1996. Creativity in context: Update to the social psychology of Creativity. Westview Press, Boulder, CO.

Andiliou, A., Murphy, P.K., 2010. Examining variations among researchers' and teachers' conceptualizations of creativity: A review and synthesis of contemporary research. Educ. Res. Rev. 5 (3), 201–219.

Arechar, A.A., Gächter, S., Molleman, L., 2018. Conducting interactive experiments online. Exp. Econ. 21 (1), 99–131.

Ariely, D., Gneezy, U., Loewenstein, G., Mazar, N., 2009. Large stakes and big mistakes. Rev. Econ. Stud. 76 (2), 451–469.

Arora, V., Thabane, A., Parpia, S., Calic, G., Bhandari, M., 2025. Generative artificial intelligence models outperform students on divergent and convergent thinking assessments. Sci. Rep. 15 (1), 36987.

Attanasi, G., Chessa, M., Gil-Gallen, S., Llerena, P., 2021. A survey on experimental elicitation of creativity in economics. Rev. d'Économie Ind. 174, 273–324.

Attanasi, G., Curci, Y., Llerena, P., Pinate, A., del Pino Ramos-Sosa, M., Urso, G., 2019. Looking at creativity from east to west: Risk taking and intrinsic motivation in socially and culturally diverse countries. In: Technical Report 21. University of Nice Sophia Antipolis.

Attanasi, G., Curci, Y., Llerena, P., Urso, G., 2026. Intrinsic vs. Extrinsic motivators on creative collaboration: The effect of sharing rewards. J. Econ. Behav. Organ. 245, 107461.

Barron, F., 1967. The psychology of creativity. Holt, Rinehart and Winston, New York.

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 Res. J. 34 (3), 245–260.
Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J.A., Bengio, Y., Jerbi, K., 2026. Divergent creativity in humans and large language models. Sci. Rep. 16 (1), 1279.
Bolden, B., DeLuca, C., Kukkonen, T., Roy, S., Wearing, J., 2020. Assessment of creativity in K–12 education: A scoping review. Rev. Educ. 8 (2), 343–376.
Boudreau, K., Lakhani, K.R., 2011. Field experimental evidence on sorting, incentives and creative worker performance. Harv. Bus. Sch. 11–107.
Burton, J.W., Stein, M.K., Jensen, T.B., 2020. A systematic review of algorithm aversion in augmented decision making. J. Behav. Decis. Mak. 33 (2), 220–239.
Charness, G., Grieco, D., 2019. Creativity and incentives. J. Eur. Econ. Assoc. 17 (2), 454–496.
Charness, G., Grieco, D., 2022. Creativity and ambiguity tolerance. Econom. Lett. 218, 110720.
Charness, G., Grieco, D., 2023. Creativity and corporate culture. Econ. J. 133 (653), 1846–1870.
Charness, G., Grieco, D., 2026. Creativity and AI. Econ. J. http://dx.doi.org/10.1093/ej/ueag015.
Cseh, G.M., Jeffries, K.K., 2019. A scattered CAT: A critical evaluation of the consensual assessment technique for creativity research. Psychol. Aesthet. Creativity, the Arts 13 (2), 159–166.
Davies, D., Jindal-Snape, D., Digby, R., Howe, A., Collier, C., Hay, P., 2014. The roles and development needs of teachers to promote creativity: A systematic review of literature. Teach. Teach. Educ. 41, 34–41.
Gneezy, U., Laske, K., Schröder, M., 2021. Creative solutions: Expertise versus crowd sourcing. Econ. Bull. 41 (4), 2580–2586.
Guilford, J.P., 1975. Creativity: A quarter century of progress. In: Perspectives in Creativity. Routledge, London, pp. 37–59.
Harris, A., Carter, M.R., 2021. Applied creativity and the arts. Curric. Perspect. 41 (1), 107–112.
Hewitt-Dundas, N., 2006. Resource and capability constraints to innovation in small and large plants. Small Bus. Econ. 26 (3), 257–277.
Hoegl, M., Gibbert, M., Mazursky, D., 2008. Financial constraints in innovation projects: When is less more? Res. Policy 37 (8), 1382–1391.
Hottenrott, H., Peters, B., 2012. Innovative capability and financing constraints for innovation: More money, more innovation? Rev. Econ. Stat. 94 (4), 1126–1142.
Kalpokas, I., 2023. Work of art in the age of its AI reproduction. Philos. Soc. Crit. 49 (2), 187–204.
Lampel, J., Lant, T., Shamsie, J., 2000. Balancing act: Learning from organizing practices in cultural industries. Organ. Sci. 11 (3), 263–269.
Lee, H.K., 2022. Rethinking creativity: Creative industries, AI and everyday creativity. Media, Cult. Soc. 44 (3), 601–612.
Li, D., Raymond, L., Bergman, P., 2026. Hiring as exploration. Rev. Econ. Stud. 93 (2), 1200–1240.
Magni, F., Park, J., Chao, M.M., 2024. Humans as creativity gatekeepers: Are we biased against AI creativity? J. Bus. Psychol. 39 (3), 643–656.
Marrone, R., Cropley, D.H., Wang, Z., 2023. Automatic assessment of mathematical creativity using natural language processing. Creativity Res. J. 35 (4), 661–676.
Murro, P., Peruzzi, V., 2026. Credit constraints and open innovation strategies. J. Econ. Behav. Organ. 245, 107543.
Oksanen, A., Cvetkovic, A., Akin, N., Latikka, R., Bergdahl, J., Chen, Y., Savela, N., 2023. Artificial intelligence in fine arts: A systematic review of empirical research. Comput. Hum. Behav.: Artif. Humans 5, 100004.
OpenAI, 2024a. Chat and file retention policies in ChatGPT. Available at: https://help.openai.com/en/articles/8983778-chat-and-file-retention-policies-in-chatgpt. (Accessed 23 January 2026).
OpenAI, 2024b. Memory FAQ. Available at: https://help.openai.com/en/articles/8590148-memory-faq. (Accessed 23 January 2026).
OpenAI, 2024c. Temporary chat FAQ. Available at: https://help.openai.com/en/articles/8914046-temporary-chat-faq. (Accessed 23 January 2026).
Organisciak, P., Acar, S., Dumas, D., Berthiaume, K., 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. Think. Ski. Creativity 49, 101356.
Patston, T.J., Cropley, D.H., Marrone, R.L., Kaufman, J.C., 2018. Teacher implicit beliefs of creativity: Is there an arts bias? Teach. Teach. Educ. 75, 366–374.
Roberts, D.L., Candi, M., 2024. Artificial intelligence and innovation management: Charting the evolving landscape. Technovation 136, 103081.
Runco, M.A., Jaeger, G.J., 2012. The standard definition of creativity. Creativity Res. J. 24 (1), 92–96.
Torrance, E.P., 1974. The Torrance tests of creative thinking. Personnel Press, Lexington, MA.