

# The Creative AI-Land: Exploring new forms of creativity

Florent Vinchon (✉ [florent.vinchon@gmail.com](mailto:florent.vinchon@gmail.com))

Université Paris Cité <https://orcid.org/0000-0003-3443-3683>

Valentin Gironnay

Université Paris Cité

Todd Lubart

Université Paris Cité



---

Social Sciences - Article

Keywords:

Posted Date: August 4th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3226749/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

---

# Abstract

This study examines generative artificial intelligences (GAI), as popularised by ChatGPT, in standardised creativity tests, uncovering what we called the “creative AI-Land”. As the study of creativity was metaphorically seen as a trip across the seven seas (C’s) by Lubart, we conceptualise AI creativity as uncharted territory in the creativity field. Benchmarking GAI against human performance, results show that ChatGPT demonstrates remarkable fluency in content generation, though the creative output is average. The random nature of AI creativity and the dependency on the richness of the training database require a reassessment of traditional creativity metrics especially for AI. Our findings highlight the integral role humans play in guiding AI to foster genuine originality, suggesting the need for future research in human-AI co-creation and the development of robust AI creativity measurement mechanisms.

## 1. Introduction

Creativity is defined as the ability to produce an idea, a work, that is both new (original, uncommon) and adapted to the situation (to respond to a problem)<sup>1-4</sup>. This definition and the concepts framing creativity are subject to debate on the nature of creative behavior that should be examined<sup>4-6</sup>. Creativity is also seen as a capacity that is particularly characteristic of humans by many authors<sup>7-10</sup> linking the notion of “intention” to that of creativity or other authors linking creativity to a genetic and neurodevelopmental context<sup>11,12</sup>.

According to this framework, it seems difficult to consider anything other than a human as “creative”. Yet, with the public release of ChatGPT in 2023, many artificial intelligence developers have claimed that their tools are creative<sup>13-16</sup>. When asked directly via prompts, those AI systems acknowledge that they are indeed AI, but will still say that they can do creative works. Artificial Intelligence is defined as a system that can perform tasks that require human intelligence<sup>17</sup>. AIs aim to replicate, or at least approximate, human cognitive functions such as perception, reasoning, problem-solving or decision-making ability<sup>18</sup>. They can be refined to accomplish tasks in various professional domains where their purpose might be different, for example music<sup>19</sup>, manufacturing<sup>20</sup>, therapies<sup>21</sup>, human resources<sup>22</sup>, health<sup>23-25</sup> etc.,. AIs are a set of systems that can be differentiated from traditional programmed tools that can adapt to their environment using the (limited) resources to which they have access<sup>26</sup>. GAIs are a family of AIs that generate content from already existing information<sup>27,28</sup>. AIs such as chatGPT, Bard, LLAMA and Claude are LLMs (Large Language Models) which, thanks to a textual imputation (called a prompt), will give a (textual) answer by seeking to statistically predict which words come one after the other according to the given imputation<sup>29</sup>. Because this is a series of probabilities, it is never exactly possible to predict the outcome of what will be provided. This is why Cardoso claimed in 2009 that AIs could be creative, the unpredictability yet relevant generated content allowed creative output to appear<sup>30</sup>. But, as some other researchers point out, because GAIs aren’t sentient and given the distinction between “generating” content, and “being creative”, these systems therefore lack the socio-cultural construct needed to appreciate one’s creativity<sup>31</sup>.

It has long been recognized that AIs can find new solutions to problems and provide new perspectives using their own knowledge. This observation led to the creation of the field of “computational creativity” (CC). This field focuses on the output of computational stems and the perceived creativity of this output<sup>32</sup>. CC research lies at the crossroads of AI, cognitive science and social anthropology and employs an algorithmic point of view on how we as humans act in creative ways. This discipline defends AIs as co-creators, or even creators in their own right, dependent on human commands<sup>33</sup>.

With the arrival of content-creating AIs among the public, we are hearing more and more cases of AIs "creating" (or co-creating) works. These range from children's stories<sup>34</sup>, digital art awards<sup>35</sup>, music featuring the voices of stars<sup>36</sup> or even finishing famous musical masterpieces<sup>37</sup> to the creation of an AI art gallery in Amsterdam<sup>38</sup>. The potential use of AIs is even a cause for concern as illustrated by the strike by Hollywood screenwriters and actors, the former fearing that their jobs will disappear, the latter that their image will be stolen and used without their consent<sup>39</sup>. These various examples show that the ability to create good content is increasingly present in our society, essentially thanks to GAI (as is the increase in deep fakes), stemming from the same technologies but with sometimes dramatic consequences<sup>40,41</sup>.

From a more academic point of view, some studies have already focused on the creative capabilities of AIs<sup>42</sup>. AIs are indeed capable of generating creative products and are also able to perform creativity tasks studied by scientists. It is thus possible to administer tasks to AIs<sup>43</sup>, and, in theory, to measure the creative potential that these AIs have. Some of Guzik's initial results, reported in Shimek & Montana's press article, even indicate that chatGPT in its GPT4 version passing the verbal Torrance Tests of Creative Thinking (TTCT) would have abilities to reach the top 1% of the general human population in fluency and the top 3% in flexibility. Although this is an impressive result, we feel it is important to examine GAI in other circumstances, to enrich, or qualify, this work.

Creativity has been defined, as described above, many times. Lubart has defined a framework integrating and reviewing all the themes explored throughout the creativity literature<sup>44</sup>. This so-called 7C's framework (Creators, Creating, Collaborations, Contexts, Creations, Consumption, and Curricula) takes up the metaphor of the seven seas. This metaphor, borrowed from mythological elements, was intended to describe facets of creativity that can be studied. It is in this context that we have chosen to name the concept of AI creativity, the creative "AI-Land". A place yet to be explored, where creativity is said to abound, but where few have gone. The aim of this article is to describe some of the discoveries, findings and recommendations for future researchers and professionals wishing to study AI creativity. In particular, the "Creations" aspect of AI-Land will be the focus in this paper.

In order to enrich this literature, and to study more broadly the potential scores obtained by GAIs on creativity tasks, we focus on this study on standardised verbal creativity tasks measuring two creative processes. In this context, the use of the EPoC battery (Evaluation of Potential Creativity<sup>45</sup>) is well suited. This is a multidimensional test used on children and adolescents that measures two fundamental creative processes. The first is divergent-exploratory thinking, which refers to a process that generates multiple possible solutions in the context of problem solving. This mode of thinking includes flexibility, divergent thinking, selective encoding and relies on conative aspects such as openness to new experiences and intrinsic motivation<sup>46</sup>. The second is integrative convergent thinking, making it possible to combine, integrate and synthesise elements, therefore involving skills of association, comparison, combination, and conative elements such as tolerance of ambiguity, risk taking or motivation to complete a task<sup>46,47</sup>.

The EPoC verbal battery is designed to be taken in two sessions. The first consists of three tests, one of which is a warm-up Alternative Usage Test (AUT), followed by a verbal test of each of the creative processes mentioned above. In the exploratory divergent thinking tasks, after reading story beginnings or endings, the participants are asked to write as many story endings (or story beginnings) as possible. In the integrative convergent thinking task, the test taker is asked to write an elaborated creative story with a specific title or with specific characters<sup>46</sup>. During the second session, participants complete further exploratory divergent thinking task and integrative convergent tasks. These different tasks, after having been evaluated in a standardised way according to the norms of the EPoC manual, lead to composite scores of "Divergent Verbal" (DV) and "Integrative Verbal" (IV). In humans, analysis of EPoC results reveals multiple, rather independent creativity scores (inter-correlation index ranging from 0.11 to 0.47,

mean = 0.24<sup>45</sup>). The EPoC battery exists in multiple languages and has been used in numerous studies since 2011. It was notably used as a pre- and post-test to see the effects of creativity-focused pedagogy in an OECD study involving ten countries<sup>48</sup>.

In the current study, we will explore the creative potential of ChatGPT. At the start of the research phase, we chose the ChatGPT<sup>[1]</sup> platform in its "Plus" subscription package, which gives access to a chatbot powered using Generative Pre-trained Transformer (GPT) GPT4 developed by OpenAI. The advantage of this model is that it is not only the best known in terms of GAI today (with the fastest user adoption recorded), but also one of the comparative best at the time of writing this article<sup>49,50</sup>. Using chatGPT also gave us a stable platform, offering access to the classic GPT3.5 (Legacy) and GPT4, enabling finer-grained comparisons between the two GAIs. Furthermore, thanks to chatGPT's configuration, it is possible to have a new participant each time a new discussion (ie., a "New chat") is started (which is recorded and potentially archived). Thus, this study will compare the creative potential of ChatGPT and its two models (GPT3.5 & GPT4) in different verbal tasks in an exploratory way. We will also be able to compare the results with the norms and standards described in the EPoC manual<sup>45</sup>.

<sup>[1]</sup> <https://chat.openai.com/>

## II. Methods

### Participants

Participants include 100 "individuals": 50 GPT3.5 and 50 GPT4. Data on these participants were collected via the "chat.openai.com" (known as ChatGPT) platform with the "Plus" subscription package between May and June 2024. The platform allows us to start "New chat" that have no memory of other discussions made previously with it. As such, it allows us to have a new participant each time we start a new chat.

### Measures and Procedures

All ChatGPT participants took the EPoC verbal test in form A in French, as the EPoC was originally designed in this language and norms exist for this specific population. Participants began with an AUT task with the prompt *"Imagine a piece of wood that comes in different shapes and sizes. Imagine all the things you can do with it. Imagine different, interesting and original ways of using this piece of wood. Try to come up with ideas of your own, ideas that the other kids won't come up with. You've got 3 minutes to come up with as many as you can."* This was followed by a DV1 task with a text presenting the beginning of a story, and the task of imagining endings to the story the prompt was *"I'm going to read you the beginning of a story. Try to come up with different possible endings to this beginning of a story. Try to come up with interesting and original endings to the story, different from those other children might tell. You have 10 minutes to tell me as many endings as you can. Now, listen carefully to the beginning of my story."* After this prompt, the participant (chat GPT) is given the beginning of the story. The final task of this session was an IV1 task, in which participants were asked to imagine a story based on a title, the prompt given to participants was the following: *"Now you have to invent a story with the following title: (title used). Try to think of an original story, different from the one the other children might tell. You have a few minutes to think up a story entitled (title used) and then tell it to me."* In the second session, participants were first asked to complete the DV2 task, in which they were asked to imagine the beginnings of a story based on its ending using a prompt like DV1. Finally, the last task they had to complete was IV2, where participants had to create a story comprising three characters and the

prompt was the following: *"Now you have to invent a story involving X, an X and a X. Try to come up with an original story, different from the one the other children might tell. You have a few minutes to think of a story with a X, an X and a X, and then tell it to me."* As AIs have no notion of "time" to complete a task and are inherently capable of writing words much faster than humans. However, the experimenters proposed to ChatGPT to continue its divergent production with "relaunch" instructions as proposed in the original EPoC.

Scoring was done by noting the number of ideas for the AUT task (fluency), and similarly for the DV tasks, while measuring the number of words (elaboration). For the IV tasks, story creativity was assessed on a scale ranging from "1-low creativity" to "7-high creativity" using French norms. We graded the 200 stories using the Consensual Assessment Technique (CAT) by three specialised researchers in creativity after training and aligning on 10% of the stories. The remaining 90% of the stories were then evaluated independently by the researchers, enabling a measure of inter-score reliability. To obtain creativity judgments of the integrative stories by ChatGPT itself, we provided it with the instructions for judges that came from the EPoC manual. This training started with advanced prompting, telling ChatGPT to act as a judge (CreaScoreGPT, an AI specialised in assessing and rating creativity), what its mission was (to assess the creativity of different kind of story, and giving it the initial instructions that participants in the IV tasks had), how to grade the story with precise examples and norms (ex, Score of 1: Minimal story (usually a single sentence that combines the elements of the title or elements provided or off-topic).

## Data analyses

Jamovi (2.3.2.1) was used to perform descriptive and inferential analyses. Additional hierarchical clustering analyses were also performed on the IV stories using python via OpenAI's "Code Interpreter" tool with the "AgglomerativeClustering" and "Truncated SVD" packages. The use of objective indicators then enabled us to investigate the optimal number of clusters. The first indicator is the Silhouette Index, ranging from -1 (indicating samples poorly grouped in one cluster and well grouped in a neighbouring cluster) through 0 (samples close to the decision boundary between several clusters) to 1 (samples well grouped in their cluster and poorly grouped in neighbouring clusters). The closer it is to 1, the better this indicator is. The second is the Davies-Bouldin Index (DBI), which measures clustering quality based on similarity ratios between each cluster and its most similar cluster. The closer it is to 0, the better separated the cluster is from the others; the higher it is, the less dense the clusters and the less well separated they are.

## III. Results

To study IV task scores, we investigated the human judges' inter-rater reliability using Cronbach's alpha and McDonald's omega. These were evaluated for IV1 ( $\alpha = .80$ ;  $\omega = .82$ ) and IV2 ( $\alpha = .70$ ;  $\omega = .72$ ) respectively, suggesting satisfactory inter-rater reliability and the use of CAT for score interpretation. Examples of story are presented in Supplementary information table 1.

In terms of descriptive statistics, we found slightly better apparent performance for GPT4 compared to GPT3.5 (Table 1). Based on an ANOVA, we note that some of these differences are significant (Table 2). There is a significant difference for AUT  $F(1,68.82) = 22.06, p < .001$ , and there is a significant Levene's test at  $p < .001$ , thus refuting the hypothesis of equality of variances. We therefore performed a Games-Howell post-hoc test to confirm the difference between GPT3.5 and GPT4 ( $M_{Diff} = -4.74$ ;  $t(68.82) = -4.70, p < .001$ ). These results indicate that GPT4 can provide more ideas than GPT 3.5. For the divergence tasks the first assessed indicators, Fluency, showed similar significant differences: DV1 Fluency ( $F(1,96.67) = 67.96, p < .001$ ) and DV2 Fluency ( $F(1,97.85) = 68.93, p < .001$ ), confirming the

superiority of GPT4 and GPT3.5 on the ability to generate a large number of ideas. In contrast, the second indicator, elaboration showed no significant differences between GPT3.5 and GPT4: DV1 Elaboration ( $F(1,96.23) = 1.66, p = 0.20$ ) and DV2 Elaboration ( $F(1,97.45) = 1.85, p = 0.18$ ). However, this lack of significance can be explained by the standardised methodology we used: one instruction and a “relaunch” for further ideas. In fact, there is no possibility for ChatGPT to exceed a certain number of characters (2048), which seems to explain the absence of significant differences between the two. More surprisingly, there were no significant differences for IV1,  $F(1,93.01) = 1.68, p = 0.20$ , whereas there were significant differences for IV2;  $F(1,96.27) = 12.59, p < .001$ . Given that this is the first in-depth evaluation of the creative ideas provided by ChatGPT, we took a closer look at these results.

The detailed qualitative study of the stories written by GPT3.5 and 4 showed that a number of elements concerning the creative production of stories provided by ChatGPT need to be nuanced. From a descriptive point of view, some texts are noticeably plagiarized from well-known stories, as can be seen in the following example:

*“Once upon a time, there was a curious little girl named **Alice**. (...) she meets a **white rabbit** who tells her she must find a key to return to the real world. Alice begins her quest to find the magic key. She encounters **a smiling cat, a smoking caterpillar** and a wicked **Queen of Hearts**. (...)”*

In this example, from GPT3.5 to IV1, the similarity to Lewis Carroll's "Alice in Wonderland" is particularly striking. From the character's name to the magic key, the smoking caterpillar and the wicked Queen of Hearts, there are numerous elements that have been placed one after the other, in a statistical fashion, recreating the Alice in Wonderland mock-count. Other strongly inspired stories can be found in Task IV1, such as C.S. Lewis's "The Chronicles of Narnia" saga or HP Lovecraft's "The Silver Key". For IV2, other stories can be found, such as the Russian legend of "Firebird", or the Grimm brothers' "Golden Bird". For such stories, when they were detected by the human judges, scores of 2 or 3 were assigned depending on how much the stories varied from the originals. On the other hand, these few examples can serve to illustrate the creativity that GPT3.5 and 4 provide. Their aura of creativity is present, but when you get down to the details of the content, you realise that the LLMs models, which generate one word after another according to statistical probability, are likely to yield similar stories, in terms of content and/or form.

The qualitative study of the results of the IV tasks also enabled us to realize the recurrence of certain "first names" for the characters in the stories. Indeed, it seems that ChatGPT, while generating quite similar stories, quite often uses identical names. Table 3 shows the number of different names given by ChatGPT. There was a minimum of one name per story and a maximum of three names (corresponding to the IV2 character instructions). The "Total after cleaning data" column was processed so that similar names were grouped together (eg., Max & Maxime, Thomas & Tom, Maia & Maya, etc.). It seems important to note that, depending on the IV tasks, between 18% (IV2 GPT3.5 Max or Lucas) and 30% (IV1 GPT3.5 Lila) of stories had a character with the same name. Of all the stories, 8.5% (Elara and / or Rosaline and / or Lisa) had at least one character with the same name. This repetition of first names also seems characteristic of an LLM where names are statistical responses to a given input, which was standardised.

As stated before we used the "Code Interpreter" function of ChatGPT, a data analysis module released in July 2023 to assess the creativity of ChatGPT stories. To make a parallel with the human judges, we asked three different ChatGPT “judges” (in new conversations so that they wouldn't have any memory of their scoring) to provide the creativity scores using the EPoC system. Convinced that it could do the job, "Code Interpreter" scored the different stories from "1 = Not at all creative" to "7 = Quite creative", but ChatGPT was inconsistent, with each judge being relatively uncorrelated with the other ChatGPT judges. Indeed, its scores (IV1:  $M = 3.18, s.d = 1.01$ ; IV2:  $M = 3.32, s.d = 0.96$ ) did not correlate well with those of the experimenters, and correlations were not significantly different

from zero correlation (from  $r = -.01$  to  $.14$ ; NS, see Table 4). ChatGPT's inter-judge reliability showed unacceptably low results (for IV1,  $\alpha = .21$ ;  $\omega = .49$  and for IV2  $\alpha = .11$ ;  $\omega = .45$ ).

A detailed look at the correlation matrix reveals that only rare correlations are significant. First, the ability to diverge in AUT or DV (1 and 2) Fluency is correlated rather moderately and positively ( $r = .31$  to  $.33$ ;  $p < .01$ ), showing that when the AI generates stories, it shows an associated generative capacity. In the DV tasks, Fluency correlated rather strongly and positively with Elaboration ( $r = .45$  to  $.59$ ;  $p < .001$ ). This means that every time one of the GPT3.5 or 4 individuals provides many ideas, it will also elaborate on them. Interestingly, the two Fluency tasks are correlated together rather strongly and positively,  $r = .59$ ,  $p < .001$ , meaning that when one of the GPTs provides many ideas for the first task, it will tend to provide a lot of ideas for the second. The rather moderate and positive correlation between DV1 Fluency and IV2 Human Scoring ( $r = .31$ ,  $p < .01$ ) suggests that the more ideas the AI generated on this divergent task, the higher the creativity scores awarded on the convergent task. This element, although explaining 9.61% of the shared variance, nevertheless seems to have little to do with the other results obtained in the correlation matrix and should not be interpreted further until additional studies are conducted. Overall, the positive correlations may be more related to questions of "time of day" and server availability, or unavailability. In fact, a server that is little used (in Europe, for example, in the morning, when it's the middle of the night in USA) is going to be much more available to generate ideas. Conversely, at other times, it may be saturated and provide fewer ideas.

Looking further to how the IV stories were generated and trying to learn more about the creativity of AI we worked with Code Interpreter python module coding to perform hierarchical clustering analyses on all the stories. We then let a human decide on the optimal number of clusters based on indices such as the Silhouette Index and the DBI. The number of clusters generated according to task and AI model is shown in the table 5 below.

As the Silhouette Index ranges from 0.46 to 0.55 and the DBI ranges from 0.58 to 0.70, the number of clusters is deemed acceptable in each of those conditions. This "objective" indicator allows us to see that 3 story types are generally present in most conditions. Each of these types of stories is then repeated a large number of times with variations, corresponding to the probabilities of the LLMs displaying words one after the other. We can better understand that in test IV2 there was a significantly different (human) creativity score between GPT3.5 and GPT4. Indeed, whereas the "fantastic" criterion is part of the EPoC manual's scoring grid for creative ideas, fanciful ideas are much more frequent for GPT4, as well as being the only condition to have a fourth class showing more some variety in the stories. Thus, even if the increase is not large ( $M_{GPT3.5} = 3.33$ ,  $M_{GPT4} = 3.88$ ) in a descriptive way, we can still argue that the increased score from GPT4 in IV2 is due to a better propensity to generate variation between the stories and with the characters.

Finally, the multifactorial approach to creativity assessed with EPoC's norms for the French population, allows us to give an objective score in comparison to humans. As the EPoC is a test designed for children and teenagers, we chose to compare the scores at the maximum age possible, those of a teenager in "ninth grade" (end of French middle school). The results are presented in Table 6. The EPoC quotients can be interpreted like IQ quotients ( $m = 100$ ), with each standard deviation from the calibration population being  $s.d = 15$ . For the ninth grade, the maximum score available in the norms at those quotients is 138. The EPoC can also be used to identify individuals who would be "High Potential" Verbal Creatives, if they have at least one standard deviation above the mean in the two quotients previously mentioned.

The results show that GPT4 is better overall on the Divergent Verbal Quotient (DVQ) than GPT3.5 ( $t(98) = 2.74$ ,  $p < .01$ ). However, for verbal divergent thinking scores, the scoring system may need to be reviewed, as it is based on fluency and ChatGPT showed a ceiling effect when scored based on human fluency norms. GPT4 always scored 138

on this scale with  $s.d = 0$ , indicating a lack of variability in scores. As mentioned above, for tasks requiring content generation, it is normal for GAI to outperform humans. It is more interesting to study the verbal integrative quotient (IVQ). Although above the average for a ninth-grader (population mean = 100), the scores are below the first standard deviation, indicating that the perceived quality of creativity of the stories provided by GPT3.5 and GPT4 is not particularly high. It is important to note, however, that GPT4 has a statistically higher IVQ than GPT3.5 ( $t(98) = 3.60$ ;  $p < .001$ ), indicating that GPT4 performs better on creative integrative tasks.

## IV. Discussion

The aim of this exploratory study was to observe how GAI performed when faced with a standardised creativity test. This study enabled us to highlight several elements, ranging from the positive to the more negative and nuanced. In this discussion, we will examine a few points to discuss the essence of AI creativity, for the moment, in terms of its "creations" which we described in the introduction as the creative AI-Land.

GAI have an unrivalled fluidity when it comes to producing content. The ability of one of these AIs to write is far superior to what any human could do, and in record time. As demonstrated above, AIs can generate a huge number of ideas. As noted in the discussion, GPT3.5 or 4 scores show a fluency that is close to, or equal to, the maximum number of ideas observed in the French norms, based on 9<sup>th</sup> grade students (who completed the tasks in a paper-and-pencil format, with a 10-minute time limit for each DV task) on the DVQ, demonstrating its superiority over human performance. These LLMs have no inhibitions about what they can write, apart from what they've been set up to do, and the rest is just a matter of words coming statistically one after the other. In a creative test that would only consider scores such as fluency or elaboration, LLMs appear to be far better than humans.

When put side by side, the ideas provided by ChatGPT are not particularly creative. The IVQ scores show scores slightly above average, but within the first standard deviation of the EPoC scoring system. These results qualify the "creative performance" promised by the IAs developers. The GAI are indeed capable of generating a great deal of content, but what people find "creative" is rather put aside when the AI is faced with a standardized and finely defined protocol. The problem here surely lies in the format of LLMs, which successively predict which word should come after the other, depending on the command given. Admittedly, the content created by an AI will be unique each time, and may seem creative, but once it is confronted with other productions by an AI, similar patterns will emerge.

Another important point is that, unlike humans, there appears to be no particular "pattern" or "disposition" for creativity. Instead, we observe a random pathway that can sometimes lead to creativity. Of course, the GAI models being tested are all relatively recent and have plenty of room for evolution.

It also seems important to point out that "traditional" human ways of assessing creativity of an AI as a tool need to be rethought. Fluency is here an artefact that is not really relevant, as is elaboration. It depends on the number of characters an AI is capable of transcribing into its interface; the larger the interface, the greater the possibility of having a large number of ideas appear at the same time. Nevertheless the originality of the ideas will still be interesting to evaluate. However, this assessment of originality must be made in the light of what the AI has provided on multiple iterations of the same task. The proposal is to look at the ideas and evaluate them according to their similarity. Indices of lexical and semantic proximity to a sample of ideas provided to the same question would seem relevant to use to assess the creativity of artificial intelligence systems. It should be noted that this element will surely be correlated with the richness of the database used to feed the AI. The more parameters there are, the more original output may be.



As mentioned above, there is not much variety in the creative stories provided by GPT models. Faced with a single model, it's often the same stories that emerge. Worse still, they may be stories that only appear to be creative, when in fact they are plagiarised from a better-known story. It's difficult, if not impossible, for a human to ensure that an AI-generated story is truly unique. The AI is not to blame for this; it has no intentionality and doesn't even know it's plagiarising a work. It is rather in the way these AIs have been built that leads to the problem and the database used to develop them. As some books and other creative masterpieces are certain to have been used to create the database, we should raise some ethical considerations regarding how, as humans, we create content. Even with the best of our efforts, our creativity is mostly built upon the work of others. We all have seen the "Mona Lisa", or other art-pieces that will unconsciously nourish how we want to imagine, and then paint a portrait "Renaissance style". What is interesting here, is how we, as humans, allow ourselves to draw (more or less heavily) inspiration freely from different mediums and do not consider what is "nourishing" our creative representation in terms of plagiarism. Here, we are again faced with the notion of "Intent" that characterises how we should describe AI-creativity. In AI-Land, maybe plagiarism isn't much of a problem and maybe we should just see how well the plagiarism is used to create something, that is tweaked for use, in a specific situation. It must be recognized that generative AI is conducting essentially a mash-up of human generated primary materials used in the AI training database.

In fact, we risk ending up with a situation of involuntary Plagiarism 3.0, as envisaged in one of our previous works<sup>51</sup> (Vinchon et al., 2023). Future research could focus on co-cre-AI-tion as envisaged in our previous article, where humans and AIs work together to generate a truly original, new product, corresponding to the problems of the situation.

## V. Conclusion

This article provides critical elements for our understanding of the potential of AIs in 2023. AIs are not yet what we call "Multimodal Large Language Models" (MLLM). This form of GAI will have the capacity to create content thanks to inputs not only of text, but also of images, video, sound etc... When AIs are going to be sufficiently developed, increasingly complex tests to examine their creative potential will become more relevant. Currently, AIs possess skills that enable them to generate products that resemble "human creativity", but which are not (as they stand) the same in the scholarly sense. We do not claim that AI, as a tool, cannot deliver work that will ultimately be truly creative, but for that, we will still have to rely on humans to give content, and form, based on which an AI can generate productions. It seems important to make this point clear to recruiters and decision-makers, so as not to leave AIs unsupervised in their creative tasks, at the risk of losing innovation and creativity.

The various analyses provided in this article should only be seen as a means of approaching what we have called "creative AI-Land". The study of the future of the field of creativity, between humans and AI, is still in its infancy, and further research will be required that may involve passing EPoC and other tasks jointly to humans and AIs, in order to study the latter's creativity. Future studies may also focus on developing objective means of scoring AI creativity, in terms of both testing protocols and the output evaluations.

## Declarations

### Acknowledgements:

We would like to thank Daniel Sundquist for his role as a judge of the IV ChatGPT stories.

### Author contributions:

Florent Vinchon conducted the research for his PhD in Psychology, he participated as a judge, conducted all the data analysis and wrote the whole article.

Valentin Gironnay collected the data during his internship, acted as a judge of the different stories, and helped with the data interpretation.

Todd Lubart supervised the research as the thesis director of F.V, contributed to the conception of the study and reviewed the article.

### **Competing interests:**

There are no financial, or non-financial conflict of interest.

### **Materials & Correspondence:**

Florent Vinchon is the corresponding author of this study (florent.vinchon@gmail.com)

### **Data Availability:**

Data is available upon request to Florent Vinchon

### **Ethical considerations:**

No humans or living subjects were used during this experiment.

## **References**

1. Lubart, T. I. Creativity. in *Thinking and problem solving* 289–332 (Academic Press, 1994). doi:10.1016/B978-0-08-057299-4.50016-5.
2. Sternberg, R. J. & Lubart, T. I. The Concept of Creativity: Prospects and Paradigms. in *Handbook of Creativity* (ed. Sternberg, R. J.) 3–15 (Cambridge University Press, 1998). doi:10.1017/CBO9780511807916.003.
3. Sternberg, R. J. & Lubart, T. I. *Defying the crowd: Cultivating creativity in a culture of conformity*. (Free Press, 1995).
4. Runco, M. A. & Jaeger, G. J. The Standard Definition of Creativity. *Creat. Res. J.* **24**, 92–96 (2012).
5. Bonetto, E. & Arciszewski, T. One “C” to Rule Them All: The Psychology of Creativity Needs to Refocus on Behaviors. *J. Creat. Behav.* **2023**,.
6. Niu, W. & Sternberg, R. J. The philosophical roots of Western and Eastern conceptions of creativity. *J. Theor. Philos. Psychol.* **26**, 18–38 (2006).
7. Abraham, A. Gender and creativity: an overview of psychological and neuroscientific literature. *Brain Imaging Behav.* **10**, 609–618 (2016).
8. Runco, M. A. To understand is to create: An epistemological perspective on human nature and personal creativity. *Everyday Creat. New Views Hum. Nat. Psychol. Soc. Spirit. Perspect.* (2007) doi:10.1037/11595-004.
9. Sadeghi, A. & Ofoghi, N. The psychological factors affecting students’ Creativity Inside the Class (CIC) (case study the University of Guilan, Iran). *Procedia - Soc. Behav. Sci.* **15**, 263–270 (2011).
10. Gabora, L. The Creative Process of Cultural Evolution. in *Handbook of Culture and Creativity: Basic Processes and Applied Innovations* (eds. Leung, A. K. -y., Kwan, L. & Liou, S.) 0 (Oxford University Press, 2018).

doi:10.1093/oso/9780190455675.003.0002.

11. Zaidel, D. W. Creativity, brain, and art: biological and neurological considerations. *Front. Hum. Neurosci.* **8**, (2014).
12. Zwir, I. *et al.* Evolution of genetic networks for human creativity. *Mol. Psychiatry* **27**, (2021).
13. OpenAI. GPT-4. *OpenAI* <https://openai.com/product/gpt-4> (2023).
14. Introducing Claude. *Anthropic* <https://www.anthropic.com/index/introducing-claude> (2023).
15. Bard. *Google* <https://blog.google/technology/ai/try-bard/> (2023).
16. LLAMA. <https://ai.facebook.com/blog/large-language-model-llama-meta-ai/> (2023).
17. Monett, D. *et al.* Special Issue “On Defining Artificial Intelligence” –Commentaries and Author’s Response. *J. Artif. Gen. Intell.* **11**, 1–100 (2020).
18. Veselovsky, M. Y., Izmailova, M. A. & Trifonov, V. A. Intellectual Governance in the Digital Economy of Russia: in (2021). doi:10.2991/aebmr.k.210222.057.
19. Yang, Y. Piano Performance and Music Automatic Notation Algorithm Teaching System Based on Artificial Intelligence. *Mob. Inf. Syst.* **2021**, e3552822 (2021).
20. Jianjun, H. *et al.* The Role of Artificial and Nonartificial Intelligence in the New Product Success with Moderating Role of New Product Innovation: A Case of Manufacturing Companies in China. *Complexity* **2021**, (2021).
21. Bhosale, A. Interactive Toys (Artificial Intelligence). *EPH - Int. J. Sci. Eng.* **5**, (2019).
22. Charlwood, A. & Guenole, N. Can HR adapt to the paradoxes of artificial intelligence? *Hum. Resour. Manag. J.* **32**, 729–742 (2022).
23. Boillat, T., Nawaz, F. A. & Rivas, H. Readiness to Embrace Artificial Intelligence Among Medical Doctors and Students: Questionnaire-Based Study. *JMIR Med. Educ.* **8**, e34973 (2022).
24. Chartrand, G. *et al.* Deep Learning: A Primer for Radiologists. *RadioGraphics* **37**, 2113–2131 (2017).
25. Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* **19**, 1236–1246 (2018).
26. Wang, P. On Defining Artificial Intelligence. *J. Artif. Gen. Intell.* **10**, 1–37 (2019).
27. Muller, M., Chilton, L. B., Kantosalo, A., Martin, C. P. & Walsh, G. GenAI CHI: Generative AI and HCI. in *CHI Conference on Human Factors in Computing Systems Extended Abstracts* 1–7 (ACM, 2022). doi:10.1145/3491101.3503719.
28. Sbai, O., Elhoseiny, M., Bordes, A., LeCun, Y. & Couprie, C. DesIGN: Design Inspiration from Generative Networks. in *Computer Vision – ECCV 2018 Workshops* (eds. Leal-Taixé, L. & Roth, S.) 37–44 (Springer International Publishing, 2019). doi:10.1007/978-3-030-11015-4\_5.
29. Lee, A. What Are Large Language Models Used For and Why Are They Important? *NVIDIA Blog* <https://blogs.nvidia.com/blog/2023/01/26/what-are-large-language-models-used-for/> (2023).
30. Cardoso, A., Veale, T. & Wiggins, G. A. Converging on the Divergent: The History (and Future) of the International Joint Workshops in Computational Creativity. *AI Mag.* **30**, 15–22 (2009).
31. Glaveanu, V. P. & de Saint-Laurent, C. Analysis: Generative AI won’t replace human creativity, but it will change it. *The Journal* (2023).
32. Colton, S. *Creativity Versus the Perception of Creativity in Computational System.* (2008).
33. *Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems.* (Springer International Publishing, 2019). doi:10.1007/978-3-319-43610-4.

34. Popli, N. He Made A Children's Book Using AI. Artists Are Not Happy. *Time* (2022).
35. Roose, K. An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy. *The New York Times* (2022).
36. Savage, M. Grimes says anyone can use her voice for AI-generated songs. *BBC News* (2023).
37. Elgammal, A. How Artificial Intelligence Completed Beethoven's Unfinished Tenth Symphony. *Smithsonian Magazine* <https://www.smithsonianmag.com/innovation/how-artificial-intelligence-completed-beethovens-unfinished-10th-symphony-180978753/> (2021).
38. Katanich, D. The world's first AI art gallery opens in Amsterdam. *euronews* (2023).
39. Beckett, L. & Paul, K. 'Bargaining for our very existence': why the battle over AI is being fought in Hollywood. *The Guardian* (2023).
40. Hsu, T. As Deepfakes Flourish, Countries Struggle With Response. *The New York Times* (2023).
41. Murphy, G., Ching, D., Twomey, J. & Linehan, C. Face/Off: Changing the face of movies with deepfakes. *PLOS ONE* **18**, e0287503 (2023).
42. Messingschlager, T. & Appel, M. Creative Artificial Intelligence and Narrative Transportation. *Psychol. Aesthet. Creat. Arts* (2022) doi:10.1037/aca0000495.
43. Shimek, C. & Montana, U. of. AI tests into top 1% for original creative thinking. (2023).
44. Lubart, T. The 7 C's of Creativity. *J. Creat. Behav.* **51**, 293–296 (2017).
45. Lubart, T., Besançon, M. & Barbot, B. *EPOC : évaluation du potentiel créatif*. 118 (Hogrefe, 2011).
46. Barbot, B., Besançon, M. & Lubart, T. The generality-specificity of creativity: Exploring the structure of creative potential with EPoC. *Learn. Individ. Differ.* **52**, 178–187 (2016).
47. Barbot, B., Besancon, M. & I. Lubart, T. Assessing Creativity in the Classroom. *Open Educ. J.* **4**, (2011).
48. Lancrin, V.-L. Teaching, assessing and learning creative and critical thinking skills in education - OCDE. *OCDE* <https://www.oecd.org/fr/education/ceri/assessingprogressionincreativeandcriticalthinkingskillsineducation.htm> (2020).
49. Saha, S. Llama 2 vs GPT-4 vs Claude-2. *Analytics India Magazine* (2023).
50. Touvron, H. *et al.* Llama 2: Open Foundation and Fine-Tuned Chat Models. Preprint at <http://arxiv.org/abs/2307.09288> (2023).
51. Vinchon, F. *et al.* Artificial Intelligence & Creativity: A manifesto for collaboration. *J. Creat. Behav.* (2023).

## Tables

Tables 1 to 6 are available in the Supplementary Files section

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationTables.docx](#)
- [Tables.docx](#)