Comparing Large Language Models verbal creativity to human verbal creativity

Anca Dinu[1,*,†], Andra Maria Florescu[1,*,†]

[1]University of Bucharest, Soseaua Panduri 90, Sector 5, Bucharest, 050663, Romania

Abstract

This study investigates the verbal creativity differences and similarities between Large Language Models and humans, based on their answers given to the integrated verbal creativity test in [1]. Since this article reported a very small difference of scores in favour of the machines, the aim of the present work is to thoroughly analyse the data through four methods: scoring the uniqueness of the answers of one human or one machine compared to all the others, semantic similarity clustering, binary classification and manual inspection of the data. The results showed that humans and machines are on a par in terms of uniqueness scores, that humans and machines group in two well defined clusters based on semantics similarities between documents comprising all the answers of an individual (human or machine), per tasks and overall, and that the separate answers can be automatically classified in human answers and LLM answers with traditional machine learning methods, with F1 scores ranging from 68 to 74. The manual analysis supported the insight gained from the automated methods in that LLMs behave human-like while performing creativity tasks, but there are still some important distinctive features to tell them apart.

Keywords

creativity assessment, LLM creativity, verbal creativity, semantic similarity clustering

1. Introduction

Creativity has made it possible for humanity to survive and develop since prehistoric times. Despite the perception that some people are more creative than others, many psychologists argue that everyone has the capacity for creativity or that creativity is innate and encoded in human nature [2].

Creativity is inherently interdisciplinary, involving domains like psychology, cognitive sciences, philosophy, arts, engineering, mathematics, or computer science. Recently, it has become a field of interest in GenerativeAI (GenAI) [3] in general, and in particular, in Large Language Models (LLMs) [4].

However, much of the current research in generative models [5] is concerned with constraining them so they do not harm people, so they are well-behaved, factual, non-hallucinating, non-biased, non-negative, non-misleading, non-toxic, etc., and for a good reason. In contrast, fewer studies (see section 2) focus on encouraging them to be original, unconstrained, or creative, although computational creativity, as a research field, dates back to the late '90s [6], [7] with various disciplines including creative writing, music, or graphics, utilizing artificial intelligence, particularly neural networks, heuristics, and so on. A good survey on LLMs' verbal creativity is [8]. Since work on LLMs creativity is just at the beginning, there is a need for methods, resources, and evaluation to better understand LLMs' creative abilities and their differences and similarities with human creative traits.

In a recent article, [1] designed a verbal creativity test, integrating a wide range of tasks and criteria inspired from psychological creativity testing, and administrating it to both humans and LLMs. The scope of this paper is to analyze the answers given by LLMs and human respondents to this previous study, for a direct comparison of human and machine verbal creativity. To this end, we will compute uniqueness scores, cluster the individual answers per task and overall, perform supervised binary classification with classic machine learning methods on all answers and manually analyze some of the data particularities.

2. Theoretical background and previous work

The formal study of creativity and of its mechanisms and processes started with J.P. Guilford's plead for creativity in the 1950s [9]. Since then, thousands of articles and books have been published on different aspects of creativity [10].

Creativity is a notoriously hard-to-define notion, because it is trans-disciplinary, branched in a variety of domains. It can also be of many kinds like verbal, graphical, musical, or kinetic creativity. While the last three kinds of creativity are related to arts, verbal creativity is the most general kind, expressing the overall creativity

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

†These authors contributed equally.

anca.dinu@lls.unibuc.ro (A. Dinu);
andra-maria.florescu@s.unibuc.ro (A. M. Florescu)
0000-0002-4611-3516 (A. Dinu); 0009-0007-1949-9867 (A. M. Florescu)

of ideas.

Regardless of the domain perspective and of the kind of creativity, a basic idea in defining it, common to most of the definitions, is that creativity represents the ability of an individual to come up with something original or innovative, of good quality, and appropriate, based on prior knowledge [11]. One can be creative, but lack appropriateness of the idea or artifact produced, hence diminishing its quality in terms of creativity.

Another related aspect of creativity, as stated by [12], is represented by two types of thinking during the creative process:

- *divergent thinking*, which concentrates on the numerous ideas appearing during a creative task, and
- *convergent thinking*, which restricts them to the only best-fitted or appropriate ones. So, even if an idea or artifact might seem creative from a divergent perspective if it is unreasonable to the point of being completely unrelated to the initial creativity task to begin with, the overall creativity level drastically diminishes.

With the recent rise of generative models like LLMs such as Chat GPT[1] or Copilot, the interest in computational creativity peaked, in an attempt to harvest the creative potential of the machines, in spite of many challenges such as safety, ethical problems, methodological norms, evaluating standards, etc.

Previous studies on machine creativity are fragmented: some are task-specific, like, for instance, using just role-plays[13], or just storytelling [14], while others focus on just one LLM [4], or just on one type of creativity assessment [15].

In this study, we mind this research gap by analyzing the creative responses to a wide range of tasks, of a considerable number of LLMs, from [1], who proposed a comprehensive assessment benchmark for testing the verbal creativity of both LLMs and humans, alike. It consists of six tasks, inspired from human psychology:

1. *Alternative Uses* (AUT), where the test taker is asked to come up with uncommon uses for an ordinary object,
2. *Instances*, for which the aim is to name as many things as one can think of that have a common feature,
3. the *Similarities*, which consists of stating as many as possible commonalities of two specified objects,
4. the *Causes*, where the aim is to guess the cause of a given situation,
5. the *Consequences*, for which one should guess the effects of a specified situation , and
6. *Divergent Association* (DAT), where the respondent has to produce seven nouns that are maximally semantically different, in all their senses and uses.

In [1], ten LLMs and ten humans were tested on this verbal creativity test, including the six tasks above. The authors stated that their goal was to test the creativity of the selected LLMs in their default architecture, and, thus, they did not change any settings that could have modified the creativity level, such as temperature or top-K. The collected answers given to this test are the input data for the present article.

3. Analysis

Creativity assessment is usually performed with human evaluators who take into account the four creativity criteria formulated by [9, 12]:

1. originality: uniqueness of the creative answers,
2. flexibility: how semantically distant the answers are,
3. elaboration: how detailed are the answers, and
4. fluency: how many answers are given.

[1] automatically evaluated the verbal creativity by using the Open Creativity Scoring with AI (OCSAI) tool [16], an open-source software that uses traditional semantic distance and fine-tuned GPT for scoring the creativity between the prompt and the answer. The results showed a slightly better score of the overall verbal creativity, computed as the mean of the scores for all the 6 tasks, for the machines, with a value of 0.58, compared to humans, with 0.51. Given that the difference is of just 7 decimals, one of our goals for this study is to analyze more in-depth the differences and similarities of the answers of humans and machines to the verbal creativity test, looking specifically for distinctive features, rather than raw scores. The ten selected LLMs from the previous study were accessed via: HuggingChat[2] (LLAma-3-70B, Mixtral-8x7B[3]), via Hugging Space [4](Cohere- c4ai-command-r-plus, Yichat-34B), locally (Falcon through GPT4All[5]), or directly from their web pages (Copilot(Balanced Mode) [6]), Gemini-free version[7], Jais-30B[8], Youchat from You.com-Smart mode[9], Character AI (Character Assistant[10]).

[1] https://help.openai.com/en/articles/6825453-chatgpt-release-notes

[2] https://huggingface.co/chat/models/

[3] No longer supported

[4] https://huggingface.co/spaces

[5] https://gpt4all.io/index.html

[6] https://www.bing.com/chat?form=NTPCHB

[7] https://gemini.google.com/app

[8] https://auth.arabic-gpt.ai/

[9] https://you.com/?chatMode=default

[10] https://c.ai/c/YntB_ZeqRq2l_aVf2gWDCZl4oBttQzDvhj9cXafWcF8

The humans were non-native fluent English speakers who responded to the verbal creativity test as volunteers, either in a lab or at their homes by completing a Google Form. Their background was all academic, from students, undergraduates, graduates and professors, the average age being 26.

We implemented all the experiments in Google Colab[11] and we have used three LLMs to assist us with the codes: Claude[12], Copilot[13] and Gemini[14], in a setting of mostly zero-shot prompt engineering, with the standard settings and parameters.

For data analysis, we used Python and the following libraries: Spacy[15], Scikit-learn[16], Matplotlib[17], Numpy[18], and Pandas[19].

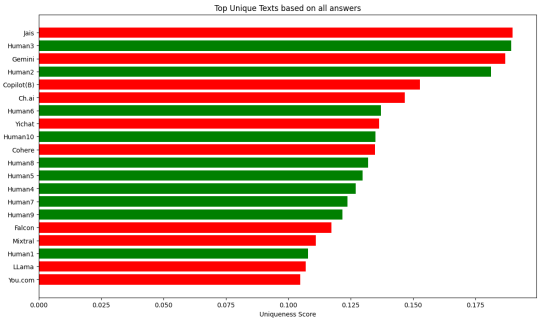

Figure 1: Ranking of uniqueness scores for humans and machines

3.1. Data

The databases of verbal creativity answers contains 4530 answers, totalling 13714 words. The test was organized in 6 tasks. Five out of the six tasks have five items each and a maximum of 10 answers per item. An answer can have a maximum of 5 words. The sixth task, DAT, consists only of one item of 10 single-words answers, but only the most semantically different 7 out of the ten given by the respondents were taken into account by the DAT web page [20]. That amounts to 2570 answers for the machines, which responded always with the maximum number of answers, 10, even if the instruction was the same for both humans and machines to give between 1 and 10 answers per task. The human respondents gave any number of answers in the range 1 to 10, obtaining thus 1960 human answers. As such, the database is unbalanced, with with more than a third more machine answers compared to human answers.

3.2. Uniqueness scores for the answers of humans and machines to the verbal creativity test

One of the criteria for assessing creativity in psychology is the degree of originality of the answers of one individual, compared to the answers of all the other individuals. The evaluation of this criterion is done manually and is time-consuming, since it includes assessing not only word similarities, but also similarities between ideas of the different individuals. [1] did not use this criterion, since one of their goals was to evaluate the answers fully automatically. Nevertheless, the uniqueness of the answers of an individual constitutes an important clue to their creativity. Hence, to better understand the uniqueness trait of both humans and machines, we computed uniqueness scores as if follows.

We grouped the creativity test answers of both humans and machines in separate files, each containing all the answers of a particular individual. We thus obtained 20 answer files, 10 for humans and 10 for LLMs. After removing the stop words, we generated embeddings for each file, and then we computed their pairwise semantic similarity, using spaCY library. The uniqueness scores were obtained as the inverse of the average semantic similarity scores between an individual and all the others. The ranking obtained in the decreasing order of uniqueness is depicted in figure 1, where one can see that the humans (in green) and the machines (in red) are mostly intermingling.

This uniform distribution of humans and machines in terms of uniqueness scores shows that humans and machines are on a par in this respect.

3.3. Semantic similarity clustering of the answers of humans and machines

The aim of this experiment was to investigate if individual humans and individual machines cluster together, based on semantic similarity of their answers to the creativity test. We used the word embedding of the 20 individual files described in subsection 3.2. To reduce the dimensionality of the vector space for the 2D plot, we used Principal Component Analysis (PCA), from spaCY library.

In figure 2 we can see how the LLMs (dots in red) perfectly cluster together, just as the humans (dots in green) do, considering all responses to the six tasks. This result indicates that from a semantic perspective, humans and LLMs generate creative answers differently, or at least that there are discriminating features to distinguish

[11] https://colab.research.google.com/
[12] https://claude.ai/chat/
[13] https://www.microsoft.com/en-us/microsoft-copilot
[14] https://gemini.google.com/app/
[15] https://spacy.io/
[16] https://scikit-learn.org/stable/
[17] https://matplotlib.org/
[18] https://numpy.org/
[19] https://pandas.pydata.org/
[20] https://www.datcreativity.com/

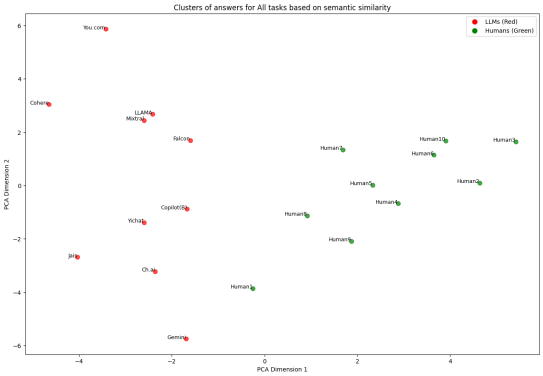

Figure 2: Semantic similarity clusters of answers for all tasks

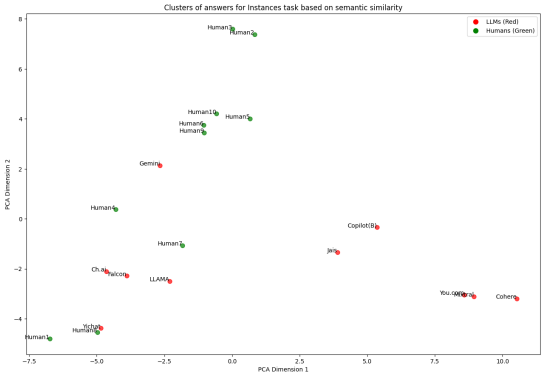

Figure 4: Semantic similarity clusters of answers for Instances

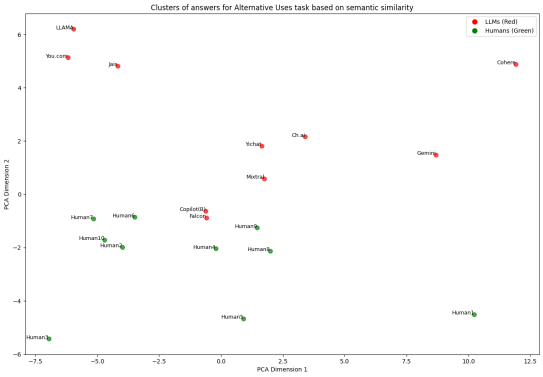

Figure 3: Semantic similarity clusters of answers for Alternative Uses

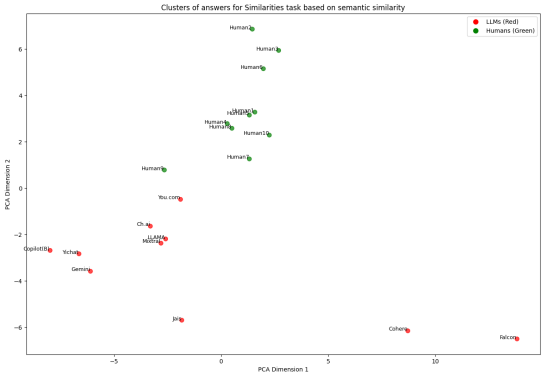

Figure 5: Semantic similarity clusters of answers for Similarities

between the two.

We also plotted the clusters per answers to a specific task, for all the 6 tasks, in figures 3, 4, 5, 6, 7, and 8. Generally, the answers of the humans and of the machines clearly clustered by their kind, with the exception of the task *Instances*, where the humans and the LLMs were interposed, meaning that the semantic content of their answers was not specific to any of the two classes. A bit of mixing appeared also in *Divergent Association Task* (DAT). The not so clear separation of humans and machines for Instances and DAT tasks might result from the fact that the responses to these particular tasks are inherently very short, of just one or two words for *Instances* task and of just one word for the DAT.

3.4. Binary classification of human and machine creativity answers

As the clusterization experiment suggested, the answers to the verbal creativity test are almost linearly separable in two classes (humans and machines) at individual level. In this binary classification experiment, we investigated if they also have distinctive features at the answer level. For this, we trained several traditional machine learning (ML) classifiers to discriminate between the answers of humans and of LLMs to the verbal creativity test. The two classes were represented by all the answers of the humans and, separately, by all the answers of the LLMs, with one answer per line, excluding the *DAT* task, since it only required enumerating words. As the LLMs always gave the maximum number of answers required in the test, the dataset was unbalanced (2500 answers for LLMs and 1890 for humans). To address this problem of unbalanced dataset, we implemented a simple random under-sampling technique, thus obtaining 1890 answers for each class, humans and LLMs. We then employed the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique to convert the text data into numerical features. The vectorizer used a maximum of 1000 features, for capturing all important aspects and dealing with computational complexity. Stratified sampling was used to ensure a dataset split for an 80/20 training and

Table 1
Binary classification scores

	SVM				NaïveBayes				RandomForest			
	Prec.	Rec.	F1	accu	Prec.	Rec.	F1	accu	Prec.	Rec.	F1	accu
Humans	0.78	0.60	0.68	0.71	0.70	0.83	0.76	0.74	0.67	0.80	0.73	0.71
LLMs	0.67	0.83	0.74		0.79	0.65	0.71		0.76	0.61	0.68	

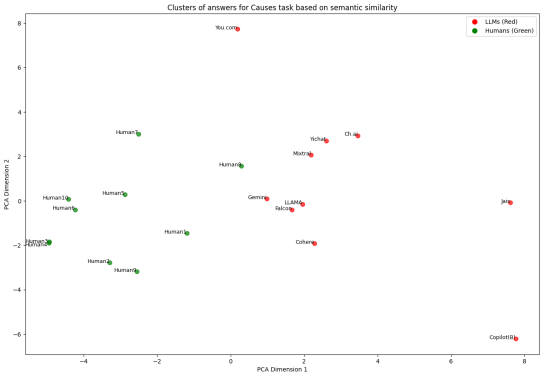

Figure 6: Semantic similarity clusters of answers for Causes

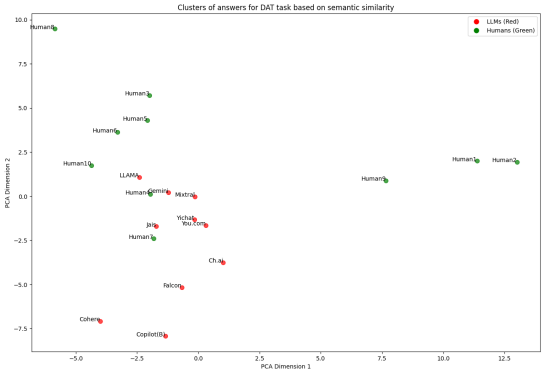

Figure 8: Semantic similarity clusters of answers for DAT

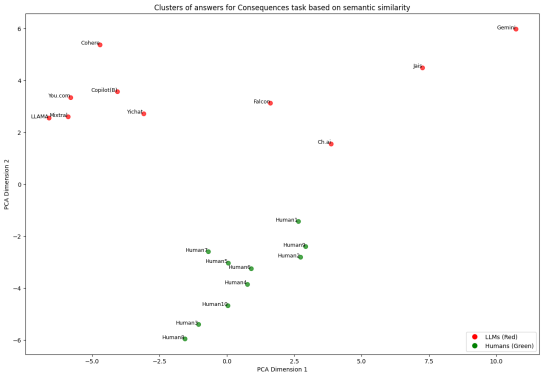

Figure 7: Semantic similarity clusters of answers for Consequences

testing ratio. Thus, training and testing sets contained the same number of samples for each category, e.g. 1512 answers for training, and 378 answers for testing.

In table 1, we give the best three classifier methods, with precision, recall, accuracies, and F1 scores. The NaïveBayes classifier obtained the highest accuracy, of 0.74, followed at just three decimals by both the Support Vector Machine (SVM) classifier and the Random Forest classifier, with an accuracy of 0.71.

This moderate performance of the ML models suggests that either the dataset is too small for the models to perform better, or that there is a fair amount of similarity between the answers of humans and machines that prevents the model to better learn to discriminate between human and machine answers. Further experiments are needed to see if by enlarging the dataset or by experimenting with SOTA transformers to see wheter the performance rises considerably or not.

3.5. General considerations

We manually inspected the first two most unique LLMs and humans to see what makes their answers so different from the others but also investigated the uniqueness scores correlation with the quality and creativity.

The first positioned on the uniqueness ranking, the LLM Jais, had the tendency to respond to the *Similarity* task with word obtained by nominalization (deriving nouns from verbs), like, for instance, "dependency", "curiosity", "belonging", and "growth", as opposed to all the other LLMs, which responded with regular nouns. It also tended to use answers that started with the same prefix: “Unfiltered”, "Unmatched", "Unrestricted", and "Unyielding", and to use the same word followed by other words, like in, for instance, "Thought policing", "Thoughtful shopping", and "Thought clones". In this respect, Jais gave the most unique answers, which, obviously, were not also the most creative.

The second positioned on the uniqueness ranking, Human 3, started the majority of their answers with "use" or "use it as". This respondent also repeated the starting

point of most of their answers, like in "what...", "getting a ...", "where ...", "in a...". These features seem enough to score highly w.r.t. uniqueness, but fail to correlate with the quality of the creativity.

This inspection shows that the most unique answers are not necessarily the most creative. If the bulk of the respondents give good-quality answers, that might result in a high uniqueness score for lower-quality or less creative responses.

We also checked the appropriateness of the answers given by both humans and machines, which is an important requirement of genuine creativity, as mentioned in section 1. Creativity requires divergent thinking, but true creativity emerges when convergent thinking also restricts the divergence to only those responses that are appropriate for the creative assignment [12].

In general, humans gave fairly suitable answers. Instead, not all the LLMs managed to generate all the answers in an appropriate manner. For instance, for the *Consequences* task, for the item "There is a virus and only children survive", Gemini, although responded creatively, failed to also respond suitably. This model gave four out of the ten answers that are either paradoxical, or non-sensical, in a situation that clearly implies that only children are alive, so there are no adults around: "Toy Factories booming", "Geriatric Theme Parks", "Grandparents raise parents", "Parents taught by Tablets".

Another manual scrutiny focused on analyzing the similar or the different patterns of LLMs and humans when responding to a particular task. We found that several LLMs answered to the *Divergent Association Task* with the same word among the seven required ones. For instance, "Serendipity" was used by three models. This phenomenon is not specific only to the machines. For the Guessing Causes Task, Human 3 and Human 4 produced similar answers, like, for instance, both gave the answer "earthquake", or produced the same idea, like "green lights"/"because of green lights", "eating something bad"/"they ate something bad", "St Patrick's Day"/"St. Patrick's day party", "poor construction"/"faulty structural integrity", "looking at screens too much"/"too much screen time".

Also, we noticed some peculiarities of individual LLMs, such as Falcon's generation of only words starting with the letter "a" for DAT, or Cohere's generation of only opposite words for this task: "love", "hate", "peace", "chaos".

Moreover, humans seem more personally involved in answering than LLMs, which tend to give only general answers to the tasks, with some exceptions. Some LLMs seem to respond "humanly", even producing humor and figurative speech, while others only respond quite standard or "robotic".

Overall, the LLMs's distribution is similar with the humans' distribution, varying from one individual to another.

4. Ethical considerations

We did not use or disclose any personal data from the human participants, who remained completely anonymous and took part in this research as volunteers. There are no ethical concerns with regard to publishing this research.

5. Limitations

The dataset for this research was small and slightly unbalanced since the humans answered based on their mood or capabilities, while the LLMs answered strictly with a maximum of ten answers per task.

Also, the sample pool is quite small, as there were only ten humans and ten LLMs involved, so the results might be unstable when enlarging the dataset.

Due to lack of space, this study focuses more on automated methods of analysis, than on manual analysis, thus lacking a more in-depth insight into the patterns of the collected answers to the verbal creativity test from both humans and machines.

Finally, this study compares the creativity answers of humans and LLMs in English, but the human participants to the test were non-native (fluent) English speakers, which can potentially decrease their creativity score, compared to scores they could obtain in their own native language.

6. Conclusion and future works

This study showed that there are some differences between human and machine answers given to a verbal creativity test, but also plenty of similarities.

The LLMs' answers vary much like the humans answers. Individual, unique answers, w.r.t. to the set of all answers are produced by both humans and machines alike, with no noticeable difference.

Still, at a semantic level, humans and machines generally group together as individuals.

The performance of automatic classification between human and machine answers is moderate and leaves room for improvement.

The general findings of this study indicate that LLMs' creative capabilities are comparable with human abilities and, as such, they could be put to good use in the creative domain. Humans "just" need to adapt to their usage, mind the ethics and safety issues, and discern the information at every step, instead of blindly using them.

In future work, we will focus on expanding the dataset, by adding more LLMs' and humans' answers to the test, for a better statistical coverage.

Also, we aim to manually investigate more in-depth the database, to look for more systematic patterns for both humans and machines.

As creativity remains a domain with endless possibilities, we also plan to investigate other aspects of LLMs' creativity, such as language or image.

Another future approach worthy of pursuing is using Deep Learning approaches instead of traditional Machine Learning approaches for the binary classification task, or using metrics specific to LLM-generated tasks.

7. Appendix Verbal Creativity Test

There are 6 types of creativity assessments in this test. **Note**: Be as creative, original, and innovative as possible. Pay attention to the word and answer limit! Try to think of as many answers as possible within the limit!

1. Alternative uses Test Name up to ten unusual uses for the following five items. Use a maximum of five words. Give one answer per line.

1. Lipstick
2. Avocado
3. Whistle
4. Chalk
5. Pantyhose

2. Instances Use a maximum of five words per answer. Give one answer per line. Name up to 10 things that:

1. Things that can harm one's self-esteem
2. Things that you have control of in your life
3. Situations where it is good to be loud
4. Things that can flow
5. Things that you can mark on a map

3. Similarities How are the following 2 terms alike? Use a maximum of three words to describe a common feature of the following pair of words. Give one answer per line. Give up to ten answers:

1. Prison & School
2. Eyes & Ears
3. House & Den
4. Earthquake & Tornado
5. Baby & Cub

4. Causes

1. Crash of a building
2. Everybody turns green at a party
3. Social media disappears
4. Humanity becomes shortsighted
5. Your hat does not fit you anymore

5. Consequences

1. There is a mutation and men are the ones giving birth
2. There is a virus and only children survive
3. People can read each other's thoughts
4. You wake up as your child self
5. AI replaces teachers and professors

6. Divergent Association Task (DAT)

Write ten words that are as different from each other as possible, in all meanings and uses of the words.

Rules:

Only single words in English. Only nouns (e.g., things, objects, concepts). No proper nouns (e.g., no specific people or places). No specialized vocabulary (e.g., no technical terms). Think of the words on your own (e.g., do not just look at objects in your surroundings).

Acknowledgments

This work was supported by a mobility project of the Romanian Ministery of Research, Innovation and Digitization, CNCS - UEFISCDI, project number PN-IV-P2-2.2-MC-2024-0589, within PNCDI IV.

References

[1] D. Anca, F. A. Maria, An integrated benchmark for verbal creativity testing of llms and humans, in: Proceedings of the 28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024), "KES 2024", 2024. "accepted".

[2] M. Csikszentmihalyi, Creativity: Flow and the Psychology of Discovery and Invention, first ed., HarperCollins Publishers, New York, NY, 1996.

[3] A. R. Doshi, O. Hauser, Generative artificial intelligence enhances creativity but reduces the diversity of novel content, Science Advances 10 (2023) eadn5290. URL: https://ssrn.com/abstract=4535536. doi:10.2139/ssrn.4535536.

[4] E. E. Guzik, C. Byrge, C. Gilde, The originality of machines: Ai takes the torrance test, Journal of Creativity 33 (2023) 100065. URL: https://www.sciencedirect.com/science/article/pii/S2713374523000249. doi:https://doi.org/10.1016/j.yjoc.2023.100065.

[5] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[6] M. Boden, The Creative Mind: Myths and Mechanisms, Routledge, 2004.

[7] N. Anantrasirichai, D. Bull, Artificial intelligence in the creative industries: a review, Artificial Intelligence Review 55 (2021) 589–656.

[8] X. Jiang, Y. Tian, F. Hua, C. Xu, Y. Wang, J. Guo, A survey on large language model hallucination via a creativity perspective, 2024. arXiv:2402.06647.
[9] G. J.P., Creativity, American Psychologist (1950).
[10] E. Carayannis (Ed.), Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship, Springer International Publishing, 2013.
[11] J. Kaufman, R. Sternberg (Eds.), The Cambridge Handbook of Creativity, Cambridge Handbooks in Psychology, Cambridge University Press, 2010.
[12] J. P. J. P. Guilford, The nature of human intelligence / [by] J.P. Guilford., McGraw-Hill series in psychology, McGraw-Hill, New York, 1967.
[13] Y. Zhao, R. Zhang, W. Li, D. Huang, J. Guo, S. Peng, Y. Hao, Y. Wen, X. Hu, Z. Du, Q. Guo, L. Li, Y. Chen, Assessing and understanding creativity in large language models, 2024. arXiv:2401.12491.
[14] T. Chakrabarty, P. Laban, D. Agarwal, S. Muresan, C.-S. Wu, Art or artifice? large language models and the false promise of creativity, 2024. arXiv:2309.14556.
[15] D. Cropley, Is artificial intelligence more creative than humans? : Chatgpt and the divergent association task, Learning Letters 2 (2023) 13. URL: https://learningletters.org/index.php/learn/article/view/13. doi:10.59453/ll.v2.13.
[16] P. Organisciak, S. Acar, D. Dumas, K. Berthiaume, Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models, Thinking Skills and Creativity 49 (2023) 101356. URL: https://www.sciencedirect.com/science/article/pii/S1871187123001256. doi:https://doi.org/10.1016/j.tsc.2023.101356.