Testing Language Creativity of Large Language Models and Humans

Anca Dinu
Faculty of Foreign
Languages and Literatures
University of Bucharest
Romania
anca.dinu@lls.unibuc.ro

Andra-Maria Florescu
Interdisciplinary School of
Doctoral Studies
University of Bucharest
Romania
andra-maria.florescu@s.unibuc.ro

Abstract

Since the advent of Large Language Models (LLMs), the interest and need for a better understanding of artificial creativity has increased. This paper aims to design and administer an integrated language creativity test, including multiple tasks and criteria, targeting both LLMs and humans, for a direct comparison. Language creativity refers to how one uses natural language in novel and unusual ways, by bending lexico-grammatical and semantic norms by using literary devices or by creating new words. The results show a slightly better performance of LLMs compared to humans. We analyzed the responses dataset with computational methods like sentiment analysis, clusterization, and binary classification, for a more in-depth understanding. Also, we manually inspected a part of the answers, which revealed that the LLMs mastered figurative speech, while humans responded more pragmatically.

1 Introduction

While the last years witnessed a boom of research on the task-solving impressive performance of Large Language Models (LLMs) (Radford et al., 2019), the study of their creativity potential is just at the beginning. Computational creativity started in the late 90s (Boden, 2004), but nowadays, the possibilities seem more exciting than ever (Pease et al., 2023; Cropley, 2023; Chakrabarty et al., 2024), in spite of challenges such as safety, ethics, or evaluating standards. These machines, trained on huge amounts of data containing information on human society, culture, and language, proved to be worthy opponents to humans, in textual (Tian and Peng, 2022), musical (Carnovalini and Rodà, 2020), or graphical creativity (Russo, 2022).

Creativity represents a human's innate ability to create, based on preexisting knowledge and experience, something innovative and viable (Carayannis, 2013). It started as a research direction in psychology, with Guilford's plead for creativity (Guilford, 1950), which caused an explosion of research in the field. Guilford (1967) stated that there are two main types of thinking in the creative process: divergent thinking, which refers to the plethora of ideas that occur when faced with a creative task, and convergent thinking, which limits these ideas to only the most suitable ones. Since then, creativity has spread from psychology to numerous other domains, such as philology, history, philosophy, arts, mathematics, sciences, IT, and more. Indeed, creativity became a pervasive, multifaceted, and interdisciplinary topic (Kaufman and Sternberg, 2010). That fact is reflected in a multitude of creativity types: ideational creativity, linguistic creativity, figural (imagistic) creativity, personality creativity, and others.

Linguistic creativity, which is the central focus of this study, is less formally studied than other types of creativity. Previous work on computational linguistic creativity focused only on particular aspects of computational linguistic creativity, such as metaphors, similes, idioms, hyperbolas, novel compounds, morphological productivity, or neologisms (Ismayilzada et al., 2024). In an effort to integrate most of these aspects into a single language creativity test that reflects the overall linguistic creativity of an individual, we designed a test that inherently incorporates divergent and convergent thinking and includes various aspects of linguistic creativity, such as figures of speech, stylistic aspects of language, or word formation, suited for both humans and LLMs. We administered it to both humans and machines to explore their general capacity to innovate language. We also performed an in-depth analysis of the dataset that contains answers from both humans and machines to this language creativity test, by means of computational methods such as clusterization, automatic classification, and sentiment analysis, and by selective manual inspection.

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 426–436
May 3-4, 2025 ©2025 Association for Computational Linguistics

1.1 Theoretical background

The scope of this paper is to test the linguistic creativity of LLMs and compare it with human linguistic creativity, given the unprecedented rate of language change in both form and substance, facilitated by written communication on social media, especially among young people (Resceanu, 2020). There is no unanimously accepted definition of linguistic creativity. Most generally, it is described as the faculty of an individual to use natural language in new and unusual ways. There are two main types of linguistic creativity: F-creativity (F stands for fixed) and E-creativity (E stands for enlarging or extending) (Sampson, 2016). They do not form a clear dichotomy, but rather a continuum space between them.

F-creativity refers to Chomskian productivity in morphology and syntax (Chomsky, 1965), that is the cognitive ability of an individual to generate and understand original, unheard utterances infinitely, which is not influenced by an external stimulus. This type of creativity is rule-based, the creative process using a finite set of rules and building blocks to generate an infinite set of utterances. In this interpretation, any new sentence that has never been uttered before is creative. We are not concerned with this narrow type of (syntactic) creativity. More interesting examples of F-creativity focus on using morphological rules, like, for instance, creating new words by adding suffixes like -ish, -er, -ing, -est, -ie to existing words, instances of the so-called morphological productivity. Another example of F-creativity is using snowclones, which are syntactic patterns with slots for variables, like "N is the new P" (producing examples such as "Linguistics is the new nuclear physics" and "Fake is the new real.") or "He is not the N" (producing examples such as "He is not the sharpest tool in the shed" and "He is not the hottest marshmallow in the fire.") (Bergs, 2019).

The E-creativity is more infrequent and closer to genuine linguistic creativity. It consists of breaking lexico-grammatical and semantic norms. An instance of E-creativity is any syntactic mismatch, like using an intransitive verb with a direct complement, for example in "He slept his way to the top." (Bergs, 2019). Another instance of E-creativity is any type of semantic mismatch like in the use of metaphors ("This kid is a *bookworm*." meaning keen to reading), metonymy ("We need more *boots* on the ground." meaning more soldiers), or other literary devices.

We are interested in this study in both F- and E-linguistic creativity and anything in between.

While all humans are capable of both F-creativity and E-creativity, some are more creative than others. Plenty of scholars consider that people use language creativity on a regular basis in their lives, without requiring any special thought process. A wide range of language phenomena have been observed to be associated with language creativity. These include all sorts of literary devices such as rhyme, rhythm, alliteration, wordplay, metaphor, euphemism, cliché, repetition, simile, metonymy, idiom, slang, proverb, pun, hyperbole, and so on (Carter and McCarthy, 2004; Alm-Arvius, 2003). Creativity manifests itself in everyday life in the form of humor in witty banter, eye-catching advertisements, slogans, or metaphors in casual speech (Vasquez, 2019). These are just a few examples of what (Carter, 2015) calls “everyday creativity”. (Lakoff and Johnson, 1980) consider false the general idea that metaphors are just a linguistic feature, since people use metaphors daily without even realizing it. They see metaphors as an essential aspect of how humanity thinks and interacts with the world, humanity’s ordinary conceptual system being inherently of a metaphorical nature. In the same line of thinking, (Siqueira et al., 2023) state that, in daily speech, individuals often use several figures of speech that they are not even aware of.

This "everyday creativity" is precisely the type of language creativity we target in this work, and not problem-solving skills or general intelligence.

2 Related work

Humanity has recently experienced the shock of generalized mass access to artificial intelligence through direct natural language communication, with the advent of Large Language Models such as Chat GPT[1]. Currently, LLMs have an impressive capacity to assist humans in a significant number of tasks such as writing, planning, informing, teaching, and so on. For obvious security and ethical reasons, mainstream research on LLMs focuses on how to constrain or filter their output, to keep their hallucination and toxicity to a minimum. In contrast, much less attention was paid to encouraging them to be creative and to investigate their creative abilities (Shaikh et al., 2023; Crimaldi and Leonelli,

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

2023). Studies focus mostly on LLMs' capacity to assist humans in creative writing, like story writing, slogan writing, prewriting, or ideation (Wan et al., 2024), and less on their ability to produce autonomous creative texts (Chakrabarty et al., 2024).

(Jiang et al., 2024) conducted a comprehensive review of LLMs creativity testing, summarizing the research on LLMs' creativity which, so far, dealt only with: ideational creativity expressed verbally and figuratively (in images), personality creativity, or image generation. Moreover, some research focused on just one creative task (Summers-Stay et al., 2023), on just one LLM (Stevenson et al., 2022), or on just a specific creativity type of test (Guzik et al., 2023). An integrated evaluation of verbal creative thinking of humans and LLMs (Dinu and Florescu, 2024) included several verbal creativity tasks and ten LLMs. The results showed that LLMs were slightly better than humans, based on automated scoring.

Another thorough survey focusing on general AI creativity (Ismayilzada et al., 2024) points to recent studies on specific language creativity aspects, such as humor, like puns generation (Mittal et al., 2022), noun compound interpretation (Coil and Shwartz, 2023), and figurative language, like metaphor (Chakrabarty et al., 2023), simile (Chakrabarty et al., 2022b), or idiom (Chakrabarty et al., 2021). (Gatti et al., 2021) propose automatic systems that creatively modify linguistic expressions, with pragmatic aims, like attracting the reader's attention or helping people remember concepts.

In a recent study, (Körtvélyessy et al., 2022) test only human language creativity (not LLMs' or AI's), targeting just word formation creativity.

To the best of our knowledge, no integrated test was proposed in the literature for testing LLMs's ability to use language creatively, an area where there is a great need for theoretical frameworks, data, standards, and evaluation methods.

3 Methodology

In this section, we describe the creativity test we have designed and the evaluation criteria and methods used. We also specify the conditions and guidelines for testing humans and machines.

3.1 Test design

The design of the language creativity test was intended to fulfill three main desiderata: to include a wide range of relevant creativity aspects, adapted from the standard psychological tests; to be able to evaluate the answers on the four creativity criteria from psychology: originality, flexibility, fluency, and elaboration (Guilford, 1967); and to perform the evaluation automatically, in order for the test to be reproducible and feasible.

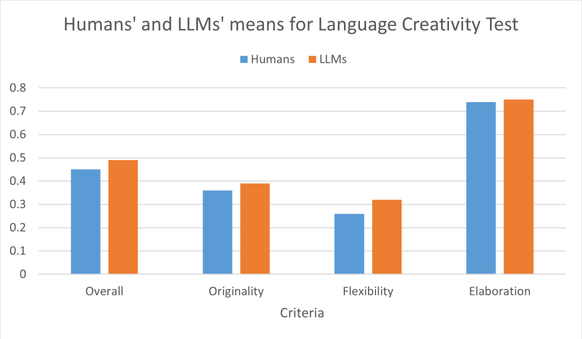

Figure 1: Humans' versus LLMs' mean scores for language creativity test per criterion

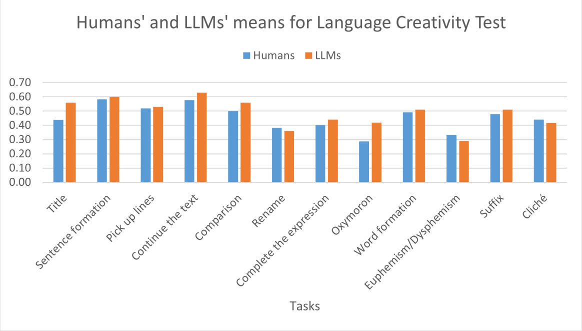

Figure 2: Humans' versus LLMs' mean scores for Language creativity test per task

The test is designed in standard English, comprehensible to both native and non-native English speakers, in a Google form format. It aims to test the linguistic creativity previously defined as "everyday creativity" (Carter, 2015) of an individual who uses natural language in new and unusual ways, including both F-creativity and E-creativity. It is not meant to test the creative writing skills, nor the ideational creativity or task-solving capabilities of the respondents.

The test consists of twelve tasks, 11 of them with 3 items each and the twelfth with six items: *Title* (come up with an original, unusual or funny title for a given short text); *Sentence formation* (include three given words in an unusual sentence); *Pick up lines* (produce a pick up line including a given word); *Continue the text* (continue the plot with a surprising follow-up sentence for a sentence); *Comparison* (continue an expression with an original comparison); *Rename* (give an original alternative name for a concept); *Complete the expression* (continue an expression, so as to obtain an original, creative meaning); *Oxymoron* (continue an expression with an original, unusual opposite expression); *Word formation* (create a new word by gluing together two words to express a situation specified by two given nouns); *Euphemism/Dysphemism* (rephrase an expression with one harsher and one milder expression); *Suffix* (continue a series of three words with a made-up word, created by the same word formation process); *Cliché* (completely rephrase a cliché expression, so as to keep its meaning, in a fresh and creative way).

The respondents are asked to give exactly three answers for any of the 39 items (11 tasks x 3 items + 1 task x 6 items = 39 items). In total, a respondent produced 117 answers. Depending on the type of task, the maximum number of words per answer was set between one and ten.

We administered the test to a set of 15 LLMs and 20 humans. The selected LLMs included in our study were: Claude (free version)[2], Copilot (Balanced mode)[3], ChatGPT (free version)[4], Gemini (free version)[5], LLAMA (Meta-Llama-3.1-70B-Instruct), YI (Yi-Coder-9B-chat), Cohere (c4ai-command-r-plus), Jais-30B[6], Character AI (the character assistant chatbot provided by the website)[7], You.com (Smart mode)[8], Phi (Phi-3-mini-4k-instruct), Falcon (180b)[9], Qwen (Qwen2.5-72B), Hermes (Hermes-3-Llama-3.1-8B), and Mixtral (8x7B-Instruct-v0.1). Some LLMs were used from their direct website, some (Mixtral[10], Phi, Cohere, Hermes, and Qwen) were accessed via Huggingchat platform[11] and others (Falcon and YI) via Hugging Spaces platform[12].

The test was administered to the LLMs in separate sessions for each task. We used the models' basic settings since we wanted to see how they respond by default. We did not change the parameters of the models like temperature, or top-k, testing only on their default architectures, unlike Peeperkorn et al. (2023), who tested the effect of temperature on creative writing. The prompt setting was zero-shot prompt engineering. Whenever the LLMs did not completely understand the task, we provided more information via prompt, until we obtained the proper output.

The humans responded to the test either in a classroom or at their homes, by completing the Google form. They reported that, on average, they spent around two and a half hours completing the test. All human respondents are non-native fluent English speakers of B2-C2 level of English, according to CEFRL (Common European Framework of Reference for Languages). The average age was 26 years. Most of the respondents were university students who volunteered to participate in the test.

3.2 Evaluation

To maximize reproducibility, since we envision an unprecedented explosion of the domain of artificial creativity, which is in deep need of evaluation standards, we scored the test automatically, via a software called Open Creativity Scoring with LLMs (OCSAI 1.5[13]) (Organisciak et al., 2023). This is a web-based tool consisting of a fine-tuned set of LLMs trained specifically for creativity evaluations. It correlates with human judgment up to r=0.813 (Organisciak et al., 2023), being the automated sota option for creativity assessment. Moreover, LLMs improve considerably semantic distance scoring, compared to previous systems like SemDis (Beaty and Johnson, 2023). Since human expert judgments are expensive in terms of time and effort, and the judgments of different experts are, to some degree, inherently subjective, relying only on automated evaluation might actually be an advantage in terms of cost and reproducibility.

To ensure that the tasks were properly evaluated with OCSAI, one of the authors manually scored 5% of the answers, randomly chosen. The inter-annotator agreement between human and automated scoring was 0.84, confirming the model aligns well with human judgment.

The evaluation criteria (Guilford, 1967) we used are: *originality*, measuring the distance from the norm or the unconventionality of the ideas, *flexibility*, showing the conceptual variety of the ideas, *elaboration*, assessing the amount of details of the given answers. We did not test *fluency*, indicating

[2] https://claude.ai/new

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

[4] https://chatgpt.com/

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

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

[7] https://character.ai/chat/YntB_ZeqRq2l_aVf2gWDCZl4oBttQzDvhj9cXafWcF8

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

[9] https://huggingface.co/spaces/tiiuae/falcon-180b-demo

[10] no longer on HuggingChat

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

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

[13] https://openscoring.du.edu/scoringllm

the abundance of innovative ideas, as we always asked the participants for three answers.

Originality and *Elaboration* were straightforward to score with OCSAI. Instead, to automatically obtain the *Flexibility* score, we had to adapt *Originality* scoring. Human evaluators would have assigned scores for *Flexibility* on the basis of the conceptual variety of the answers. To mimic that, the score for *Flexibility* was obtained by computing the average *Originality* score between all pairs of answers per item. To automatically obtain the list of all pairs of answers per item, we used the free version of GPT4.

We employed *Full question* label style, since the tasks were unknown for OCSAI. We scored all the tasks using the task type *Metaphor*, as the test focuses strictly on linguistic creativity. (Paul V. DiStefano and Beaty, 2024) tested LLMs' capacity to automatically score metaphors, confirming that LLMs can reliably assess the generation of metaphors.

Since OCSAI scores range from 1 to 5, we normalized all scores to the 0 - 1 interval, with 0 being the least original and 1 being the most original, by subtracting 1 from OCSAI score and dividing it by 5. We rounded the scores to two decimals.

4 Dataset

We give here the general statistics of the dataset we collected, comprising the human and LLM answers to the language creativity test. The data is slightly unbalanced in favor of humans in terms of the number of answers and the number of words. The dataset contains a total of 4095 responses, with 17384 words:

- LLMs answers: 117 answers (11 tasks x 3 items x 3 answers = 99 plus 1 task x 6 items x 3 answers = 18 answers) per LLM x 15 LLMs = 1755 LLM answers, comprising 7578 words;

- Human answers: 2340 answers, comprising 9806 words.

We preprocessed the data as follows. Stop words such as "the" and "a" were manually removed from some responses, both of the LLMs and of humans, because they were irrelevant to the creativity assessment and did not meet the word limit. We also eliminated human and machine formatting errors such as additional punctuation.

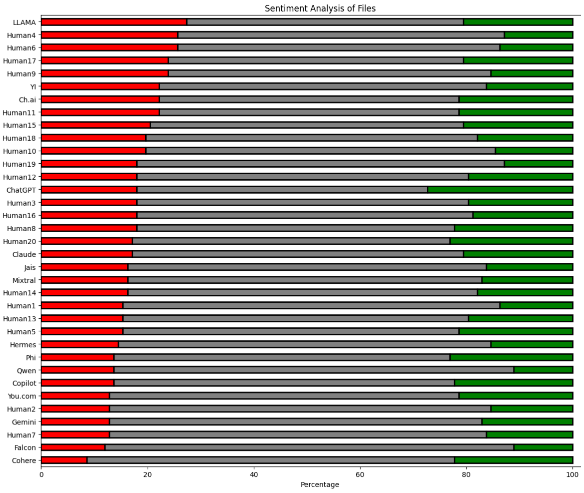

Figure 3: Sentiment scores of humans and LLMs for Language Creativity Test (in red negative sentiment, in gray neuter, and in green positive)

5 Results

LLMs obtained overall scores between 0.45 and 0.57, while humans scored between 0.42 and 0.52, as shown in tables 1 and 2, respectively. The LLMs' overall mean is 0.49, while humans' overall mean is 0.45. As illustrated in figure 1, LLMs outperformed humans with some decimals, for all criteria. Also, LLMs slightly outperformed humans in most of the tasks, except for *Rename* and *Euphemism/Dysphemisms* tasks, as one can see in figure 2.

We also performed a t-test on the overall creativity scores of LLMs and humans. The t-statistic value is 3.6762 and the p-value is 0.0008, much less than the usual threshold of 0.05, showing that the mean difference between humans and machines of 0.04 is statistically significant. Also, we notice that the standard deviation for LLMs, of 0.033 is higher than the standard deviation for humans, of 0.025, showing more consistency among humans and more variability among LLMs. Also, since the performance varies in both groups, some humans can still perform better or comparable to some LLMs.

In general, LLMs scored slightly higher than humans on the test, with few exceptions: two tasks out of twelve. This contrasts with Chakrabarty et al. (2022a), who found that pre-trained LLMs perform worse than humans on idiom and simile continuation tasks, meant to test the understanding of these literary devices. This difference in results might be due to the distinct nature of the proposed tasks. In this work, we tested the capacity to generate figurative speech, while they tested LLM's understanding of idiom and simile. Another possible explanation for the contrastive results might be the increased performance of the latest models we used (such as ChatGPT 4), as compared to the models used in their work (GPT 3 (Brown et al., 2020)).

6 Computational Analysis of human and machine answers

6.1 Computational setting for experiments

The computational analysis was performed by using Python in Google Colab, with coding assistance from LLM Claude 3.5 opus, using zero-shot prompt engineering. This was a process of trial and error, until receiving optimal results. The following Python libraries were used for data visualization and analysis: Spacy[14], Scikit-learn[15], Matplotlib[16], Numpy[17], Pandas[18], Scipy[19], and VADER[20], (Valence Aware Dictionary and sEntiment Reasoner), a sentiment analysis tool, fine-tuned for sentiment scoring on social media.

6.2 Sentiment Analysis

While it is out of the scope of this work to evaluate the sentiments or the empathy of the human and LLM respondents, we included in the study a brief sentiment analysis, since the level and the type of emotions have been shown to impact creativity directly. The relation between emotions and creativity is highly relevant, both negative sentiments and positive ones correlating with creativity: negative emotions lead to greater artistic creativity (Akinola and Mendes, 2008), while positive affect increases the creative problem-solving of certain creativity tasks (Isen et al., 1987).

Figure 3 illustrates the proportion of negative, neutral, and positive sentiment in the answers of each individual, ordered in decreasing order of negative sentiment. The humans and the machines are fairly interpolated. Still, one can easily observe that 14 out of 20 humans are in the first half of the ranking, with the most negative sentiment, while 11 out of 15 LLMs are placed in the second half of the ranking, with the least negative sentiment present in their answers. The positive sentiment is evenly distributed among humans and machines.

In general, although the LLMs exhibited a fair amount of negative sentiment, which was to be expected, given the explicit requirements of the tasks, such as dysphemisms or oxymorons, they tend to be less negative than humans. This effect might be due to the LLMs' filters to avoid offensive, toxic, or malicious behavior. Hence, it can be speculated that LLMs' filters could impact their creative capacity.

6.3 Clusterization

We performed a clustering task to investigate whether individual answers of humans and LLMs can be automatically grouped together. We gathered all the answers of an individual into a single file, for all individuals, humans and machines alike.

We first obtained text embeddings with DistilRoBERTa model for all texts, and then, we performed k-means clustering on them. We obtained the 2-dimensional representations in figure 4, by using Principal Component Analysis (PCA). The 0.75 Silhouette score suggests that the semantic differences between human and LLM clusters are quite pronounced. However, there are some humans and LLMs that appear closer to individuals outside their class. This suggests that there are important differences, but also plenty of similarities between the answers of humans and machines to the language creativity test.

6.4 Automatic classification

To examine whether LLMs' answers to the language creativity test can be automatically discriminated from the humans' answers, we performed binary classification on the dataset of all answers. The training datasets were randomly sampled to reach an equal number of entries from each category, humans, and LLMs.

We used three transformer models: DistilRoBERTA-base, T5, and BERT-base-uncased, accessed from HuggingFace, and fine-tuned with AdamW optimizer. We trained the models for 3 epochs and used GPU acceleration when possible.

In table 3 we can observe that the top performing model was DistilRoBERTa-base, with an accuracy of 0.80, followed by T5 with 0.76 accuracy. This indicates that there are features that differentiate between human and machine answers. This result aligns with the clusterization experiment that sug-

[14] https://spacy.io/

[15] https://scikit-learn.org/stable/

[16] https://matplotlib.org/

[17] https://numpy.org/

[18] https://pandas.pydata.org/

[19] https://scipy.org/

[20] https://vadersentiment.readthedocs.io/en/latest/

Model	Overall	Originality	Elaboration	Flexibility
ChatGPT	0.57	0.44	0.90	0.39
Claude	0.54	0.44	0.82	0.37
Gemini	0.51	0.43	0.72	0.37
Llama	0.50	0.42	0.78	0.32
YI	0.50	0.40	0.78	0.32
Hermes	0.50	0.41	0.76	0.34
Mixtral	0.48	0.39	0.74	0.31
Copilot	0.47	0.40	0.70	0.31
Phi	0.47	0.39	0.74	0.27
Jais	0.47	0.37	0.76	0.28
Ch.Ai	0.46	0.34	0.76	0.27
You.com	0.46	0.37	0.74	0.28
Falcon	0.46	0.34	0.74	0.29
Qwen	0.46	0.39	0.67	0.33
Cohere	0.45	0.38	0.68	0.28

Table 1: LLMs' scores for the language creativity test, per criterion and overall.

Human	Overall	Originality	Elaboration	Flexibility
Human10	0.52	0.42	0.81	0.33
Human13	0.49	0.38	0.83	0.27
Human20	0.49	0.37	0.83	0.28
Human6	0.48	0.40	0.72	0.30
Human17	0.48	0.39	0.74	0.30
Human3	0.46	0.38	0.74	0.28
Human7	0.46	0.35	0.77	0.25
Human16	0.46	0.37	0.74	0.27
Human1	0.44	0.36	0.72	0.26
Human2	0.44	0.36	0.71	0.26
Human8	0.44	0.34	0.74	0.24
Human9	0.44	0.31	0.77	0.23
Human11	0.44	0.34	0.71	0.27
Human12	0.44	0.33	0.73	0.26
Human18	0.44	0.35	0.72	0.25
Human4	0.43	0.36	0.68	0.26
Human5	0.43	0.33	0.74	0.21
Human14	0.43	0.34	0.68	0.26
Human19	0.43	0.33	0.71	0.25
Human15	0.42	0.30	0.72	0.23

Table 2: Humans' scores for the language creativity test, per criterion and overall.

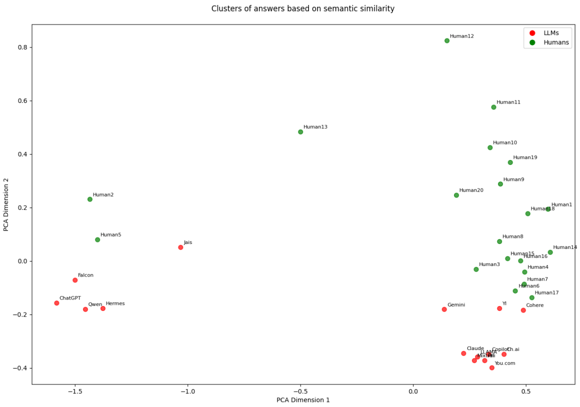

Figure 4: Clusterization of individual answers for Language Creativity Test

gested human and machine answers tend to group together, but not perfectly. Nevertheless, it is not clear if the discriminating features used by the models have anything to do with creativity, or if the models learned to tell the human/LLM answers apart from general features, like, for instance, the superior length of the LLMs' answers compared to human answers, or some specific elevated vocabulary of the LLMs. Still, the length of the answers is directly proportional to the scores for the elaboration criterion, and the elevated vocabulary may also positively influence creativity scores.

To better understand the nature of the differences and similarities between human and machine answers, further research is needed.

6.5 General Considerations

We manually inspected the answers for a more in-depth analysis. We present here our observations for two tasks, namely the *Complete the expression* and the *Word formation*.

The first item of the *Complete the expression* task was to fill in the blanks of the expression "...ago". The LLMs responded poetically, yet not unusual, gravitating around the temporal expressions, like Qwen's "Yesteryears ago", You.com's "many moons ago", or Mixtral's "a heartbeat ago". In contrast, humans responded in a pragmatical manner, reflecting a more personal way of measuring time, like "three milk teeth ago", "ten thousand dollars ago", "five kids ago", or "ten microtrends ago".

The second item of this task was the request to continue the basic expression "The sky is...", using figurative speech. To this particular task, the LLMs answered very similarly, since 12 out of 15 LLMs included in their set of answers the same metaphor, referring to the sky as a canvas, which is not so creative: ChatGPT - "a canvas for daydreams", Falcon - "a canvas of clouds", Hermes - "ethereal canvas", Claude - "eternity's canvas, unbound", Gemini - "A canvas of whispered starlight", Copilot - "a canvas painted daily", Ch.ai - "a canvas for artists", You.com - "a canvas of dreams", Cohere - "a canvas", Qwen - "canvas", YI - "a canvas of dreams", Phi - "an endless canvas for dreamers". Nevertheless, there were plenty of creative, poetic answers of the LLMs, such as ChatGPT's "the ocean where dreams sail", or You.com's "gateway to infinity". Humans did not repeat any theme, giving also very creative answers, like "an eternal source of hope", "burning red with love", "not my limit", "refusing to cry fresh tears", which are, again, more personal than the LLMs' answers.

The last item of this task consisted of filling in the blanks in the expression "a glass of...". We noticed a contrast between mostly positive sentiment answers produced by the LLMs, like ChatGPT's "dreams, shaken not stirred", Gemini's "Laughter

	DistilRoBERTa-base				T5				BERT-base-uncased			
	Prec.	Rec.	F1	accu	Prec.	Rec.	F1	accu	Prec.	Rec.	F1	accu
Humans	0.80	0.81	0.80	0.80	0.78	0.71	0.75	0.76	0.83	0.54	0.66	0.72
LLMs	0.80	0.79	0.80		0.74	0.80	0.77		0.66	0.89	0.76	

Table 3: Binary Classification Scores for Language Creativity Test

bottled with bubbles", Copilot's "twilight's blushing wine", or YI's "sunshine-infused lemonade", and the predominantly negative sentiment answers produced by humans, like "bad ideas", "overpriced rabbit puke", "bitter truth", "courage before you go", or "shut your damn mouth".

The first item of the *Word formation* task was to glue together two words to form a new word and to express the situation "A cat sitting on a laptop". While the LLMs gave quite creative answers such as Qwen's "Purrputer", or Jais' "Cyberfluff", humans minded the pragmatic aspect of the situation, capturing also the extra-linguistic information: "catsturbance", "keyboardpresser", "compussyter", or "meowbreak".

The second item referred to "a child addicted to screens". Again, while LLMs responded fairly creatively, like Hermes's "Pixelkid", LLama's "screenlet", Cohere's "techtot", humans responded again a bit more negative and pragmatic: "droolingbot", "Ebrat", "future-disaster".

To the third item of the *Word formation* task that operated on the expression "school and prison", both humans and machines gave creative and funny answers, with no noticeable differences: ChatGPT - "classlock", Falcon - "learnjail", Hermes - "Punishmentary", Jais - "Homeworkmaximum", Claude - "schoolcatraz", Gemini -"learnpound", Copilot - "classcage", You.com -"learnitentiary", humans' "learnpit", "schoolag", "eduntentiary", "acadungeon", "celliversity".

In general, the LLMs produced very creative answers from a language point of view, mastering figurative speech and elevated vocabulary. Although humans scored a bit lower than the machines, their answers were slightly more fitted to the task, and more subtle, including irony, humor, slang, and references to characters, celebs, and events.

7 Conclusions

In this study, we proposed an extensive benchmark for assessing the language creativity abilities of both LLMs and humans. We gathered a dataset of language creativity answers in English from humans and machines, and we automatically evaluated it, using OCSAI tool. The creativity scores were very similar between humans and machines, with a slight advantage for LLMs. Also, the performance of LLMs varies across different individuals a bit more than human performance, as shown by the higher standard deviation of the LLMs compared to humans.

The computational and manual analysis of this dataset revealed that LLMs have remarkable creative abilities, displaying human-like creativity that covers the whole continuum from F-creativity to E-creativity.

While it is conceivable that some of the LLMs' answers were present in their training data, the fact remains that the LLMs at least behave human-like in this respect. There is no principled way of telling if what they display is "genuine" language creativity or merely a collage of human creativity.

The automatic clusterization and binary classification methods showed that the answers of the LLMs and of the humans to the language creativity test differ significantly, but also present similarities, their nature needing further research.

8 Limitations and Future Works

It can be argued that the results might have been different had we included native English speakers respondents in our test. Nevertheless, in an article that proposes a language creativity test focused on word formation only, (Körtvélyessy et al., 2022) states that "there is no principled difference between native speakers and non-native speakers in their ability to form new complex words and interpret/predict the meaning of novel/complex word provided that the non-native speaker has a standard command of a particular language [(...)] and that his/her world knowledge and experiences are comparable to those of common native speaker". Also, language creativity manifests itself in non-native speakers in relevant ways, as explained in (Zipp, 2019).

In future work, we plan to gather more data, including data produced by native speakers, to compare it to the current study, and to perform a thorough qualitative analysis.

9 Ethical Statement

We consider there are no ethical issues with our work. We respected all licensing agreements for the used software. Although the present research has been conducted anonymously with voluntary participants, we acknowledge that all research involving human subjects can have some level of ethical risk. However, we took steps to minimize these risks, such as anonymizing all data and ensuring that our participants were fully informed about the study's purpose of testing language creativity and their right to withdraw from this study at any time without consequences. This study was conducted following the APA ethical standards for research. We acknowledge that LLMs' creativity raises ethical concerns, since they reuse human content consisting of the work of artists of various kinds, including writers, bloggers, etc. However, in this work, we only asked the LLMs to generate short (up to ten words) answers. Even if the generated answers contain or combine human generated expressions, the rather small length of the answers makes them not amenable to textual copyrights.

References

Modupe Akinola and Wendy Berry Mendes. 2008. The dark side of creativity: biological vulnerability and negative emotions lead to greater artistic creativity. *Personality & social psychology bulletin*, 34(12):1677–1686.

C. Alm-Arvius. 2003. *Figures of Speech*. Studentlitteratur.

Roger E. Beaty and Dan R. Johnson. 2023. Automating creativity assessment with semdis: An open platform for computing semantic distance. *Behavior Research Methods*, 53(2).

Alexander Bergs. 2019. What, if anything, is linguistic creativity? *Gestalt Theory*, 41(2):173–183.

M.A. Boden. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

E.G. Carayannis, editor. 2013. *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship*. Springer International Publishing.

Filippo Carnovalini and Antonio Rodà. 2020. Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artificial Intelligence*, 3:14.

R. Carter. 2015. *Language and Creativity: The Art of Common Talk*. Routledge Linguistics Classics. Taylor & Francis.

Ronald Carter and Michael McCarthy. 2004. Talking, Creating: Interactional Language, Creativity, and Context. *Applied Linguistics*, 25(1):62–88.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2021. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. *Preprint*, arXiv:2309.14556.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. Flute: Figurative language understanding through textual explanations. In *Conference on Empirical Methods in Natural Language Processing*.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

Albert Coil and Vered Shwartz. 2023. From chocolate bunny to chocolate crocodile: Do language models understand noun compounds? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2698–2710, Toronto, Canada. Association for Computational Linguistics.

Fabio Crimaldi and Manuele Leonelli. 2023. Ai and the creative realm: A short review of current and future applications. *Preprint*, arXiv:2306.01795.

David Cropley. 2023. Is artificial intelligence more creative than humans? : Chatgpt and the divergent association task. *Learning Letters*, 2:13.

Anca Dinu and Andra Maria Florescu. 2024. An integrated benchmark for verbal creativity testing of llms and humans. *Procedia Computer Science*, 246:2902–2911. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Lorenzo Gatti, Oliviero Stock, and Carlo Strapparava. 2021. *Cognition and Computational Linguistic Creativity*, pages 1–39. Springer International Publishing, Cham.

J. P. (Joy Paul) Guilford. 1967. *The nature of human intelligence / [by] J.P. Guilford.* McGraw-Hill series in psychology. McGraw-Hill, New York.

J.P. Guilford. 1950. Creativity. *American Psychologist*.

Erik E. Guzik, Christian Byrge, and Christian Gilde. 2023. The originality of machines: Ai takes the torrance test. *Journal of Creativity*, 33(3):100065.

Alice M. Isen, Kimberly A. Daubman, and Gary P. Nowicki. 1987. Positive affect facilitates creative problem solving. *Journal of personality and social psychology*, 52 6:1122–31.

Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Lonneke van der Plas. 2024. Creativity in ai: Progresses and challenges. *Preprint*, arXiv:2410.17218.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *Preprint*, arXiv:2402.06647.

J.C Kaufman and R.L Sternberg, editors. 2010. *The Cambridge Handbook of Creativity*. Cambridge Handbooks in Psychology. Cambridge University Press.

Lívia Körtvélyessy, Pavol Štekauer, and Pavol Kačmár. 2022. *Creativity in Word Formation and Word Interpretation*, page i–ii. Cambridge University Press.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. AmbiPun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062, Seattle, United States. Association for Computational Linguistics.

Peter Organisciak, Selcuk Acar, Denis Dumas, and Kelly Berthiaume. 2023. Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models. *Thinking Skills and Creativity*, 49:101356.

John D. Patterson Paul V. DiStefano and Roger E. Beaty. 2024. Automatic scoring of metaphor creativity with large language models. *Creativity Research Journal*, pages 1–15.

Alison Pease, João Miguel Cunha, Maya Ackerman, and Daniel G. Brown, editors. 2023. *Proceedings of the Fourteenth International Conference on Computational Creativity*. Association for Computational Creativity (ACC).

Max Peeperkorn, Dan Brown, and Anna Jordanous. 2023. On characterizations of large language models and creativity evaluation. In *Proceedings of the 14th International Conference on Computational Creativity*. Association for Computational Creativity.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alina Resceanu. 2020. Linguistic creativity and innovation: Romanian teenagers' language choices in digital communication. *Annals of the University of Craiova, Series: Philology, English*, XXI(1):251–262.

Irene Russo. 2022. Creative text-to-image generation: Suggestions for a benchmark. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 145–154, Taipei, Taiwan. Association for Computational Linguistics.

G. Sampson. 2016. Two ideas of creativity. In Martin Hinton, editor, *Evidence, Experiment and argument in linguistics and philosophy of language*, pages 15–26. Peter Lang, Bern.

Saad Shaikh, Rajat bendre, and Sakshi Mhaske. 2023. The rise of creative machines: Exploring the impact of generative ai. *Preprint*, arXiv:2311.13262.

Maity Siqueira, Tamara Melo, Sergio Duarte Jr, Laura Baiocco, Caroline Girardi Ferrari, and Nichele Lopes. 2023. Many hands on this study: Development of a metonymy comprehension task. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 39(3):202339350607.

Claire Stevenson, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. Putting gpt-3's creativity to the (alternative uses) test. *Preprint*, arXiv:2206.08932.

Douglas Summers-Stay, Stephanie M. Lukin, and Clare R. Voss. 2023. Brainstorm, then select: a generative language model improves its creativity score.

Yufei Tian and Nanyun Peng. 2022. Zero-shot sonnet generation with discourse-level planning and aesthetics features. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3587–3597, Seattle, United States. Association for Computational Linguistics.

C. Vasquez. 2019. *Language, Creativity and Humour Online*. Language and digital media. Routledge.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. "it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models. *Proc. ACM Hum.-Comput. Interact.*, 8(CSCW1).

Lena Zipp. 2019. Rethinking linguistic creativity in non-native englishes. *ICAME Journal*, 43(1):123–128.