Are Large Language Models More or Less Creative than Humans? – Blog des Kulturwissenschaftlichen Instituts Essen (KWI-Blog)

blog.kulturwissenschaften.de/llms-and-creativity

29.07.2024

Are Large Language Models More or Less Creative than Humans? Von: Louise Röska-Hardy

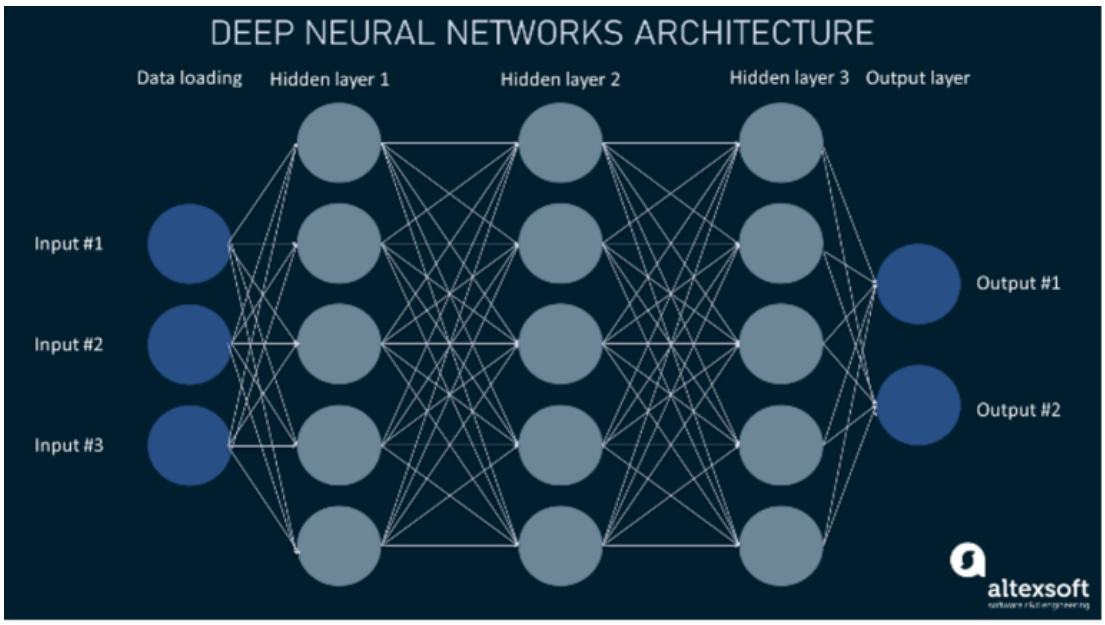
What does it mean to be creative? Until recently, creativity was considered a uniquely human capacity that sets us apart from nonhuman animals and machines. But with the development of Large Language Models (LLMs) like GPT-4, LLaMA 2 or Gemini that produce humanlike texts, conversation, poetry and program code, the question arises whether such large-scale Generative Artificial Intelligence systems are creative, too.

Creativity is hard to define, but two criteria are widely acknowledged. Creative ideas, acts or products are novel (new, original) and valuable (appropriate, useful, effective, suitable, fitting) in a given context. Novelty is necessary for creativity, but not sufficient, because there is such a thing as “original nonsense” (Kant). Something can be novel in that it has never been expressed or produced before, but neither appropriate nor valuable in a context. Novelty as a result of chance, ignorance or delusion is not considered an expression of creativity, even though it may be original. “Creativity” is the capacity to produce something new and appropriate in a given context, whereas an idea or product is “creative” if it satisfies the dual criteria in context.

Creativity is a multi-faceted psychological construct. All measures of creative capacity are indirect, since creativity is not directly observable. It must be inferred from observations that allow researchers to test, study and measure aspects of creative ability. The Alternative Uses Task (AUT) and the Torrance Tests of Creative Thinking (TTCT) are standard tests used to measure human creative thinking potential.[1] The AUT involves thinking of alternative uses for everyday objects. The TTCT measures creative thinking ability by assessing subconstructs of creativity. It is given to children and young people to identify creative potential. A high score indicates creative potential, making creative behavior more likely, but does not ensure creative achievement. Neither the AUT nor the TTCT purport to measure creativity as such, although they are often so understood. They measure divergent thinking, a proxy for creative ability.[2] Divergent thinking ability is not identical with creativity, since it only involves coming up with multiple ideas, whereas creativity involves producing an idea that is both novel and fitting. This requires evaluating new ideas to select the most fitting in context. Moreover, the AUT and the TTCT do not require a test taker to evaluate any of the ideas for usefulness. Thus, the AUT and TTCT are not measures of creativity as such for they only track one aspect of creativity, whereas creativity requires producing ideas that are both new and suitable in context. Recent claims about LLM creativity on the basis of AUT or TTCT tests often overlook this point.

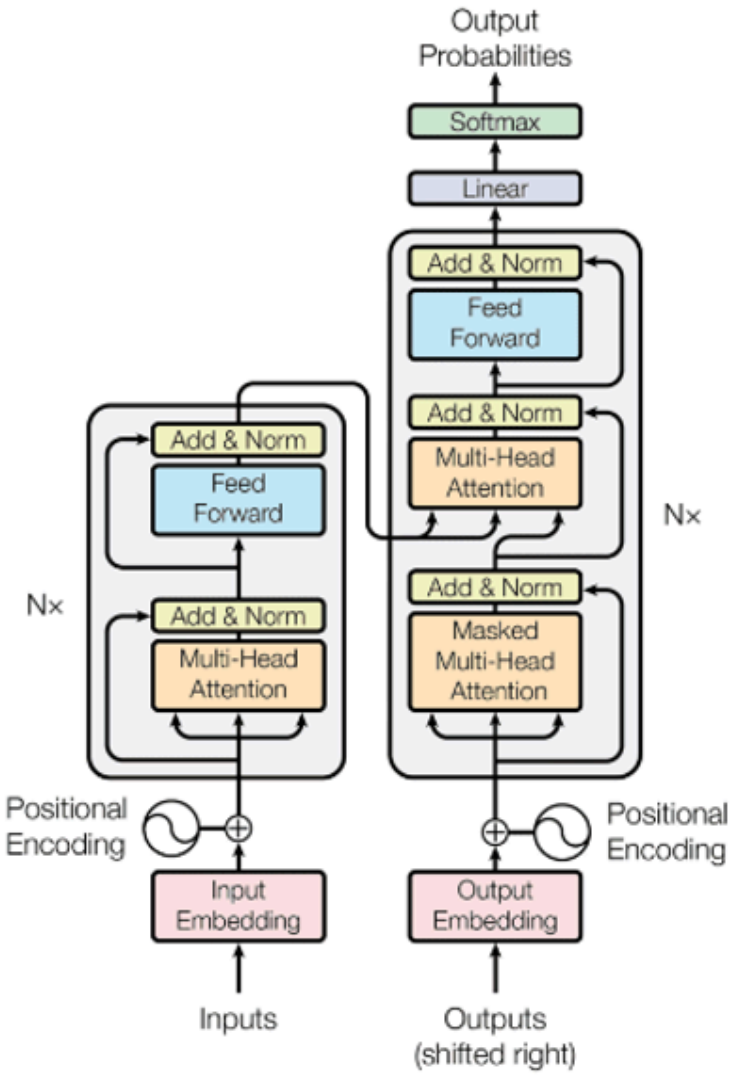
Some studies find that LLMs' answers on the AUT and TTCT are not qualitatively different from human-generated answers.[3] However, the process LLMs use to generate answers differs importantly from the way humans arrive at their responses. Briefly, experiments in psychology and neuroscience indicate that the human creative process comprises two main phases – idea generation and idea evaluation.[4] The process of idea generation involves recognizing a problem, need or challenge, searching semantic and episodic memory systems, information gathering and retrieval processes like associative thinking. Research indicates that regions of the brain's subconscious Default Network are then recruited to generate ideas.[5] After a period of incubation, a novel insight, solution or idea seems to occur spontaneously. Evaluating the idea engages the brain's conscious Central Executive Network to assess the idea for originality and appropriateness.

In contrast, LLMs are massive, neural network-based models that use numbers, vectors or arrays of numbers and statistical correlations in multi-level computations in neural networks to generate responses to input. Neural networks are machine learning algorithms, inspired by the structure and function of the human brain. They consist of multiple layers of interconnected nodes or artificial "neurons" in an input layer, one or more "hidden" layers and an output layer. The input is passed position-wise through the network layer by layer. Each layer receives input from the previous layer, performs a computation and sends the result to the next layer.



LLMs consist of layers of neural networks in a transformer architecture.[6] The transformer receives an input, encodes it and then decodes it to produce an output prediction, e. g., the statistically most probable next token or word in a text sequence. GPT-3 (Generative Pretrained Transformer-3) has 96 transformer layers; GPT-4 is reported to have 120. These steps are repeated multiple times for each hidden layer of the transformer, e. g., 96 times for GPT-3. With each layer the model takes into account increasingly complex relationships and patterns to better predict the next word.

Before an LLM like that powering ChatGPT can answer AUT prompts like “Come up with as original and creative uses for pants as you can”, it must be trained. The first step is pretraining a foundational neural network. LLMs are pretrained on large human language corpora, sourced from the web, digitalized books, Wikipedia, etc. The LLM is fed vast amounts of text with the goal of having it predict the next word or a hidden part of an input sequence in self-supervised learning. The model “learns” statistical patterns, the relationships between words, phrases and sentences, linguistic structures and semantic properties in order to “understand” and generate humanlike language. It also “learns” to distinguish words like BANK based on context. The result is a statistical model of how the words and phrases in its dataset, comprising billions to trillions of words, are related. Basic neural networks are then trained with supervised fine-tuning to perform specific tasks like translation, coding, content creation or question answering. After multiple iterations, pretraining and fine-tuning enable the LLM to use transformer architecture to process, predict and generate text content in response to input from human users, e. g., asking ChatGPT a question.



”Attention is all you need” (Vaswani et al. 2017), fig. 1

LLMs generate responses by first breaking up text input into tokens – units like words, subwords and characters, a process called tokenization, and issuing each a unique numerical index and a numerical representation, known as an embedding. Embeddings transform the token sequence into a vector sequence. The tokenizer sends each piece of text to a known token in the LLM’s library, transforming words into lists of numbers or vectors, mathematical representations that the model can process. High-dimensional vector embeddings represent the token’s syntactic and semantic meaning in context and how it relates to other tokens in the model’s vocabulary. Words with similar meanings are represented by vectors with similar numbers. Vectors for positional encoding are added to embeddings to retain the word order of the text and the embeddings are passed into the transformer, composed of an encoder and a decoder, to predict the next token in the sequence.

The transformer is a stack of multiple transformer blocks. Each block consists of a multi-head self-attention mechanism and a position-wise feedforward layer, a neural network in which information flows in one direction from input to output. Instead of processing data serially, the attention mechanism enables the model to process all the parts of the input simultaneously in order to identify dependencies between words and to determine which parts are most important, performing parallel computation.

In self-attention the transformer analyzes how each element in a sequence relates to all other elements. Transformers use self-attention mechanisms in the encoder, in the decoder, and again in encoder-decoder multi-head attention of the decoder. By “attending” to itself, the transformer identifies and weighs the importance of different parts of the input sequence. This allows the model to focus on relevant information and to capture relationships between different elements in the sequence, regardless of their distance from one another.

Multi-head self-attention builds on self-attention. The multi-head attention mechanism consists of multiple self-attention layers running in parallel. Multi-head attention lets the attention mechanism focus on several parts of the input sequence simultaneously to capture the full context of the input sequence, regardless of length. A sequence is divided into several “heads”, each of which focuses on a different part and captures different features of the sequence. The multi-head attention mechanism calculates for each part an attention score or weight, which determines how much “attention” each part should receive relative to the others. The results from all the self-attention layers are then combined to form the output of the multi-head attention mechanism. This output is then passed through a position-wise feed-forward neural network to a normalization layer, which transforms the inputs into a standard distribution.

The output of the encoder stack is then fed into the decoder. The decoder layers mirror the encoder layers with several additional layers for generating the next token and transforming the resulting vectors into output token probabilities, which are used to pick the next word. The repetition of these processing steps produces humanlike texts in response to user prompts, creating the impression that the LLM understands human language, whereas an LLM only “understands” numbers. Nonetheless, the output of LLMs is often judged by users to be creative. Does this mean that LLMs are actually creative and perhaps even more creative than humans?

The answer lies in the character of LLMs. LLMs are machine learning algorithms that operate on multi-dimensional vector embeddings to generate a statistical prediction of the next token in a text sequence. LLMs’ responses to prompts are the result of mathematical calculations on vectors in multiple layers of neural networks, repeated multiple times. The model doesn’t “consider” content. Talk of LLMs “learning”, “seeing” or “understanding” anthropomorphizes the models, because, taken literally, they cannot do any of these things. Hence understanding the structure of LLMs and how they work is essential to deciding if LLMs are creative and whether they are more or less creative than humans.

The creative process in humans starts with recognizing a problem, need or challenge, leading to a goal-directed search for pertinent information as well as the activation of memory systems to find a solution or generate an idea. In contrast, processing in LLMs only takes place when human users enter prompts. As machine learning algorithms, LLMs have no agency. However, under the guidance of users’ goal-directed prompts, LLMs can generate outputs that human users judge to be creative. Humans must make the assessment, because LLMs are unable to determine whether outputs satisfy the criteria of novelty and appropriateness. In particular, LLMs are unable to determine whether the many ideas they generate are of value in a given context.

In sum, LLMs produce potentially creative output, which human users assess with respect to originality and appropriateness to determine when an output is creative. LLMs are “creative” in the sense that they can produce outputs that humans judge to be novel and appropriate in a context and thus creative. But LLMs are less creative than humans. Human users’ prompting is required to actualize LLMs’ potential for creative output and to determine when it is creative. LLMs lack the impetus to create. So far, creativity across the board remains a quintessentially human capacity.

References

1. Clapham, M. M. (2011) Testing/Measurement/Assessment, in: Runco, M. A. & Pritzker, S. R. (eds.) *Encyclopedia of Creativity* vol. 2, (3rd ed., pp. 458-464) San Diego: Academic Press.
2. Runco , Mark A.· (2019):· *Divergent thinking*, in: *J. C. Kaufman* & *R. J. Sternberg* (*eds.*)*,* The *Cambridge handbook* of *creativity*(2nd *ed.,* pp. 224–254) Cambridge: Cambridge University Press.
3. Guzik, E. E. ,Byrge, C. & Gilde, C. (2023): The originality of machines: AI takes the Torrance test. *Journal of Creativity* 33(4), 100065; Haase, J. & Hanel, P. H. (2023). Artificial Muses: Generative artificial Intelligence chatbots have risen to human-level creativity. https://arXiv.2303.12003. (accessed on May 21, 2024).
4. Beaty, R. E. et al. (2016): Creative Cognition and brain network dynamics. *Trends in Cognitive Sciences* 20(2), 87-95.
5. Röska-Hardy, L. (2023): Das Default-Netzwerk: Die Quelle des Neuen? in: Friedrich Jaeger & Sabine Voßkamp, *Wie kommt das Neue in die Welt? Kreativität und Innovation interdisziplinär*. Berlin: Metzler pp.247-263.
6. Transformer architecture is a type of deep learning model introduced in „Attention is All You Need“ by Vaswani et al. (2017). https://arxiv.org/abs/1706.03762 (accessed on May 21, 2024).

SUGGESTED CITATION: Ruska-Hardy, Louise: Are Large Language Models More or Less Creative than Humans?, in: KWI-BLOG, [https://blog.kulturwissenschaften.de/llms-and-creativity/], 29.07.2024

DOI: https://doi.org/10.37189/kwi-blog/20240729-0830