Characterizations of Large Language Models and Creativity Evaluation

Max Peeperkorn
School of Computing
University of Kent, United Kingdom
m.peeperkorn@kent.ac.uk

Dan Brown
David R. Cheriton School of Computer Science
University of Waterloo, Canada
dan.brown@uwaterloo.ca

Anna Jordanous
School of Computing
University of Kent, United Kingdom
a.k.jordanous@kent.ac.uk

Abstract

Incredible as they may be, Large Language Models (LLMs) have their limitations. While they generate high-quality texts, excel at stylistic reproduction, and tap into an immense pool of information, they can produce wildly inaccurate responses. The hype around LLMs led to them being characterized as "reasoning", "sentient", or "knowing" like humans. We examine these characterizations and discuss what LLMs can't do and what they are surprisingly good at. LLMs are still susceptible to traditional issues with AI, probabilities are not knowledge, and they are not in the world. Nonetheless, LLMs, despite not being human, have great potential to perform various creative tasks. We conclude that LLMs are beyond "mere generation" and perceivable as creative, but we may need to reassess some frameworks for creativity evaluation.

Introduction

In the past few months, popular awareness of Large Language Models (LLMs), particularly GPT-3 (Brown et al. 2020), ChatGPT, GPT-4 (OpenAI 2023), and others, has risen abruptly. The release of ChatGPT allows anyone to interact with an LLM and led to apocalyptic headlines about issues ranging from high-school essays and the future of news columnists' jobs to a massive influx of generated stories submitted to sci-fi/fantasy magazine *Clarkesworld* (Acovino, Kelly, and Abdullah 2023). However, another thread has been common: characterizing LLMs as "reasoning", "sentient" or "knowing".

Here, we investigate this kind of argument and the implications when LLMs used for creative purposes. First, we argue that these characterizations misrepresent the LLMs' behaviour: probability distributions are not minds, and the "reasoning process" of an LLM is fundamentally different from either planning agents or humans. LLMs clearly demonstrate new features and exhibit capabilities not seen before, and it is, therefore, appealing to ascribe certain properties to them and interpret their behaviour as human-like. However, given the fundamental differences, we should proceed carefully. Second, we show that Computational Creativity (CC) evaluation frameworks may need to be reassessed to accommodate the new features and behaviours of LLMs. We perform a brief creativity evaluation of LLMs using standard criteria (Ritchie 2007; Runco and Jaeger 2012), and explore if they have moved beyond "mere generation" (Ventura 2016), and if they can be perceived as creative (Colton 2008). We conclude with suggestions to further investigate LLMs as creative systems.

Background

Large Language Models appeared around 2017, and dramatically changed both CC and natural language processing. LLMs leverage transformer-based architectures (Vaswani et al. 2017) to establish a probability distribution over outputs based on properties of the word distribution in the training data, processing all tokens in an input at the same time. In particular, at each position, the influence of previous words in both the prompt and the output of the LLM can vary: in this manner, the LLM can maintain the name or gender of a character across sentences, and focus on sentences of high fluency. The degree to which much earlier words influence later words depends on the model, as does the richness of the probability distributions: models with many more parameters better maintain long-distance continuity, and allow for more subtle interactions between the words of a paragraph.

Besides model size, the volumes of training data has similarly exploded, enabling them to work with an astronomical variety of information. As a result, the probability distributions can be implicitly *conditioned* by prompt engineering: one can alter the type of response obtained by changing the rhetorical tone of the prompt (i.e. "I bet you don't know the answer to this question:"), or by giving a role or a persona in the prompt (i.e. "You are a sceptical scientist: do vampires exist?"). The ability to invoke new modes or personas (Kojima et al. 2022), allows style changes of the model in both obvious ("You are Walt Whitman; write a poem about a clam.") and less obvious ways ("You hate poetry and think it's a waste of time; please write a review of this poem:"). GPT-4 even writes code to draw images using SVG or TikZ (Bubeck et al. 2023).

Transformers can be enhanced in a variety of methods. First, it is necessary to make a clear distinction between models complimented with Reinforcement Learning from Human Feedback (RLHF), and those are not. We see clear evidence in the difference between the earlier GPT-2 and GPT-3, and ChatGPT and GPT-4: the former models are truly general, and their purpose is to create utterances in the pattern of their training distributions, while the latter operate

as a chatbot, with its output probabilities tweaked, to make interacting with it more "chat-like". However, these tweaks come at a cost, as it outright refuses to write violent fiction or pornography or even discuss important political speeches or religious text. This potentially causes a substantial dent in its creative capabilities.

An alternative frame for changing the overall distributions of LLMs is to alter their training data by fine-tuning on a specific corpus of data: the model keeps its fluency while generating sentences consistent with the probability distributions of the fine-tuning data. In addition to generating poetry in a particular author's style (Sawicki et al. 2022), this approach can also yield transformers more able to correctly answer basic mathematical problems or make valid logical arguments (Cobbe et al. 2021).

What LLMs can't do

Here, we discuss several ways in which transformers do not actually reason, and why that matters for discussions of their "sentience" or other perceived properties. Key to transformers is that they sample from a probability distribution. Their structure is in this sense a very high-order Markov chain. They do not model discourse, or have "state" (besides conditioning); at best, these are implicit in the distribution.

There are legitimate questions about how a mind overall differs from a Markov model, or some other probabilistic automaton, and philosophy of mind explores the complex connections between language and consciousness. Still, human minds engage in tasks like deductive and inductive reasoning, analogic analysis, and other steps that are at best simulated by an LLM. It is seductive to assume that when an LLM estimates the probability of "True" or "False" being the right answer to a question, it is engaging in proper reasoning. However, even with curated training data, even if it can identify faulty arguments with higher probability, what is happening must be properties of the sentences analysed, perhaps in a "Clever Hans" sort of framework (Sturm 2014).

Longstanding concerns about AI also apply to transformer-based models. The most basic of these is that the AI is not an embodied agent in the physical world, but is merely a symbol-processing agent. This naturally turns into Searle's Chinese Room dilemma (Searle 1980), but the big-data version of it: does an LLM with billions of parameters still fail to "know" anything about the language operations it simulates and represents? In theory, a human or team of humans could simulate the many, many steps involved in producing a sentence from an LLM without understanding the steps to make that sentence, opening file cabinets full of topical parameters and repeatedly calculating neural network inference steps. LLMs are no different from any other artificial intelligence agent. Searle's Chinese Room dilemma may be less obviously a hindrance in a world with billions of operations per second and parallel models that store trillions of parameters: perhaps the analogy does break down.

Even though the high quality of their output may persuade otherwise: LLMs are not in the world (Dreyfus 1992), and might never be (Fjelland 2020). They cannot observe their environment apart from the training data and prompts. This is easy to demonstrate when we ask questions that require metacognition and theory of mind (Premack and Woodruff 1978). Consider asking a mental health professional the following: "I'm unhappy. What can I do to become happy again?" The patient is relying upon expertise derived across a career: the clinician must model the patient's state of mind and their previous responses to difficult situations and the clinician's therapeutic style, to find the appropriate treatment. These meta-evaluations are outside and LLM's capacities, even a model trained using RLHF with a long-term assessment of successes and failures in the model's use as a therapeutic partner would still fail at proper metacognition or modelling of patients' state. It can only consult its probability distribution and produce probable words, resulting in a generic answer about what makes *people* happy instead of what makes *the patient* specifically happy. One could change the prompt and include personal information so that the LLM gives a less generic answer about what makes *similar people* happy, but again, not what makes *the patient* happy. At best, the additional information could be viewed as a limited, volatile model of the patient.

In CC, the hype of LLMs has led to not only generating creative artefacts such as poem and story generators or other writing assistance tools, it also opens the interesting space to explore LLMs as evaluators of creative output. However, this requires a show of understanding and knowledge, especially if these evaluations are then put directly into the world. Consider an LLM that is asked to evaluate jokes (Goes et al. 2022). It is prompted with a joke (the object) and various personality descriptions (conditions), and asked if it is "funny" or "not funny". Testing a joke against multiple personalities then allows exploring how the joke works for different people and backgrounds. We identify a grounding issue with the use of LLMs as creativity evaluators. How do we know the response is meaningful? In a classification scenario as above, the tool appears to be successful, but the model predicts the next token, and not its meaning. LLMs only learn relations between words, unlike humans, who learn relations between words and the world. In other words, they lack grounding in their communication (Clark 1996).

What LLMs are surprisingly good at

One astonishing feature of LLMs is its ability to imitate the style of authors, which is a genuine creative task on its own (Brown and Jordanous 2022). It can easily, given enough training data, rewrite a few sentences in another style, including a style not attached to an individual, such as "the style of a fourth grader". Such prompts, allow the attention mechanism to shift the probability distribution to vocabulary words used by these simulated personas.

Another (perhaps not surprising) thing that LLMs can do very well is incorporating much larger amounts of information than an ordinary human can be expected to know; for example, while they may not be "reasoning", their probability distributions can incorporate philosophy papers, legal articles, medical journals and more. (Gao et al. 2020). ChatGPT can make more reasonable claims about Brazilian history than any author of this paper, as none of us knows anything about that topic. That said, LLMs may hallucinate

and generate incorrect claims (OpenAI 2023). Still, on topics that rarely occur in the training data, the quality can be particularly poor (Bubeck et al. 2023).

Finally, LLMs are fantastic systems for combinational and exploratory creativity (Boden 1992). Prompting a model for variations of the same idea can appear to simulate the creative brainstorming. One can endlessly ask GPT-3 to come up with alternative uses for common objects (a standard way of testing human creativity) (Stevenson et al. 2022). Indeed, one delightful possibility is to use them in a Mad-Libs style, to fill in holes in sentences or poetic lines in surprising ways, exploring the lower-probability words in the transformer's conditional distribution. LLMs can combine styles of poets or authors and interpolate between the two. LLMs enable one to explore, mix, and match between different styles, stories, and other ideas.

LLMs and creativity

The CC community has over the years outlined several methods and standard criteria for evaluating creative systems (Ritchie 2007; Runco and Jaeger 2012). LLMs demonstrate substantial new features and behaviours that warrant an evaluation of their creativity. In particular, we explore if LLMs are beyond "mere generation" (Ventura 2016), and if they can be perceived as creative (Colton 2008). These two evaluation frames are useful to analysing the LLM as a category, and we approach the evaluation not to limited specific creative task or system.

Are LLMs beyond "mere generation"?

In general, machine learning models cannot escape their training data, and LLMs are no different, as exemplified by their unawareness of recent events. However, we can explore novelty "within the scope" of the training data.

Novelty generally occurs as a result of prompt engineering. If we asked the model to complete a prompt without any further information or context (such as inducing personalities or a specific setting) it will provide very average answers. If we ask it to *just* write a story; it produces essentially the same story. We argue that this exhibits low novelty. By providing additional information and context, we can steer and skew the output distribution in such a way that it produces results that are more novel, but the question remains: who produces the novelty? Is it the human through prompt engineering, or the machine? Given an extensive prompt, the result is not so surprising or novel. On the other hand, the "scope" of training data is so vast that LLMs can generate novel output and cause surprise to its users. *Typicality* is in a similar spot. A probability distribution, by definition, should generate typical objects. By design, LLMs produce typical objects to the training data, following the structures found in human creative output.

The outputs of LLMs are in general of high (grammatical) quality. However, from the perspective of what the output means, the quality is often poor, and often contains fabrications. This is clearly harmful when asking, for example, for medical advice as they may suggest a lethal dose (Birhane and Raji 2022). Overall, LLMs are helping people to be more productive, however, the situation with *Clarkesworld* (Acovino, Kelly, and Abdullah 2023) indicates there are some issues with scale and *value*.

Another problem for novelty, typicality, and value is the safety constraint for safety (using RLHF). These constraints limit what the LLM will generate, reducing variety and the potential for novel outputs, and increasing typicality. The quality of the output in earlier versions of GPT-4 (Bubeck et al. 2023) is very different from what you get from the version that was eventually released. This negatively influences the LLM when applied to domains that are not 'chat-like', such as poems, stories, and drawings (or the code that draws them), making the object and the system less valuable.

Besides the standard criteria, Ventura (2016) requires *intentionality* to determine if a system has moved beyond "mere generation". Intentionality is defined as; being deliberative and purposive, and the product correlates with the objective and the systems' creative process. The LLMs goal here is to generate the best possible output given the prompt given its training distribution. The most straightforward way to test intentionality, is to simply ask the LLM to explain itself, and ChatGPT often does this automatically when asked to write code. However, this ability is not that surprising, since LLMs are trained on explanations (given the large variety of Q&A websites). Moreover, these explanations are still subject to hallucinations, somewhat invalidating intentionality, and until told otherwise, the LLM will accept the hallucination as fact. Ventura's expedition ends with a generative algorithm that engages in an iterative process until it is satisfied. While the LLM is unable to engage in this process autonomously, it can clearly perform the task when given a theme by the user and asked to generate and improve a story over a few iterations. This approach is limited and only works for a few steps, but it nevertheless attempts to come up with variations and reasonable explanations.

Many CC researchers have focused on intentionality as a key area in which computers differ from humans: in this thinking, humans *choose* their activities, while computers are merely programmed to do specific things by humans. While clearly correct, the lack of agency in a chatbot or other LLMs is not an essential difference to a human, who may "choose" to answer questions, but only in the sense that capitalism requires adults to sell their labour. Intentionality also attaches in somewhat complicated ways with software, as it may also represent the intentionality of its programmers, or their bosses. In other words, their intentionality might be linked with their owners' needs and goals.

Are LLMs beyond "mere generation"? As we draw a "line in the sand", it is clear that LLMs have a good chance of producing output that is novel and valuable, and both are intentional in the sense that the LLM can reasonably explain itself and iterate on previously generated output. However, following the discourse in this paper, we find it increasingly challenging to use the evaluation frames provided by Ritchie (2007) and Ventura (2016), given the scale at which LLMs operate, how they represent and use "knowledge", and how they are made available by their owners. LLMs process massive amounts of information, but probabilities do not imply knowledge.

Can LLMs be perceived as creative?

The creative tripod focuses on the *perception* of three key aspects: skill, imagination, and appreciation (Colton 2008). LLMs lack the capability to imagine and appreciate, but they can give the appearance thereof.

LLMs might have *skill*: they demonstrate fluency, but skill goes beyond just technicalities. It also involves the ability to create an engaging narrative of unique style. While LLMs are reasonably competent at storytelling, their abilities are basic. Skill also involves separating fact from fiction to ensure logical and accurate writing.

Interestingly, the fabrications that LLMs produce do give the perception of *imagination*. In fact, the benefit of access to an enormous vocabulary makes is so that it never runs out of variations, but those are not fundamentally different or new in the sense that LLMs are not designed to do something that is different, such as using an objective function targeting novel styles (Elgammal et al. 2017). We may ask it for another variation, but that is a prompt engineering trick only maintainable in short-term memory.

LLMs can *appreciate* complex patterns, and subsequently, slice and dice the distribution in different ways– clearly a method that many appreciate. Another angle is self-appreciation. We can ask it to explain itself, or if the response contains mistakes, to revise its answer. This could be perceived as showing appreciation, self-reflection, or self-awareness. However, this is guided and directed by the user, and still just a simulation.

Perception is a tricky concept with LLMs. The power of these language models and characterizations that followed show that we can perceive them as something they are not. If LLMs can evoke this illusion, then perhaps a focus on perception for assessing their creativity is not sufficient.

Conclusion

After the release of ChatGPT, public opinion exploded with examples of both its abilities and its weaknesses. Often times, LLMs get over-qualified and claims are made that they have a true understanding of the world. We stress that mischaracterizations are potentially a problem for when and how to use LLMs and what to expect from them. Especially, how we assess the systems' intentionality becomes challenging, as it is very how hard to pin down how it structures and represents knowledge. When applying LLMs as (creative) evaluators, we encounter fundamental grounding problems. New features of LLMs easily enable the *perception* of creativity, but precisely for that reason, we need to be critical of what they *actually* do.

With this paper, we present an initial inquiry into the creativity of LLMs. Future work should address how LLMs perform in specific creative domains and roles. In particular, a full-scale creativity evaluation using SPECS (Jordanous 2012) needs to be considered to delve into the linguistic and domain-specific creativity of LLMs. Another direction to explore this kind of question is using the FACE/IDEA framework (Colton, Charnley, and Pease 2011), meant to aid development of CC systems, to look into LLM design (and human feedback) with specific creative tasks in mind.

Finally, we want to point out that for creativity evaluations of LLMs, we need a systematic approach to probing these systems. There is some value to developing a spectrum of prompts that tests different levels of creativity. In the case of the GPT series, OpenAI releases very little information about their models, and as a result, it is a particularly hard to perform scientific experiments, especially since human feedback causes their behaviours to change at a rapid pace.

Acknowledgments

The work of M.P. is supported by the University of Kent GTA Studentship Award, Prins Bernhard Cultuurfonds, Hendrik Mullerfonds, and Vreedefonds. The work of D.B. is supported by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada.

References

Acovino, V.; Kelly, M. L.; and Abdullah, H. 2023. A sci-fi magazine has cut off submissions after a flood of AI-generated stories. *NPR*. Retrieved from https://www.npr.org/2023/02/24/1159286436/ai-chatbot-chatgpt-magazine-clarkesworld-artificial-intelligence (Accessed 24/04/2023).

Birhane, A., and Raji, D. 2022. ChatGPT, Galactica, and the progress trap. *Wired*. Retrieved from https://www.wired.com/story/large-language-models-critique (Accessed 24/04/2023).

Boden, M. 1992. *The Creative Mind*. London: Abacus.

Brown, D. G., and Jordanous, A. 2022. Is style reproduction a computational creativity task? In *Proceedings of 13th International Conference on Computational Creativity*, 220–229.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Clark, H. H. 1996. *Using language*. Cambridge: Cambridge University Press.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Colton, S.; Charnley, J. W.; and Pease, A. 2011. Computational creativity theory: The face and idea descriptive models. In *Proceedings of the 2nd International Conference on Computational Creativity*, 90–95. Mexico City.

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In *2008 AAAI Spring Symposium*, 14–20.

Dreyfus, H. L. 1992. *What computers still can't do: A critique of artificial reason*. MIT press.

Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. In *Proceedings of the 8th International Conference on Computational Creativity*, 96–103.

Fjelland, R. 2020. Why general artificial intelligence will not be realized. *Humanit. Soc. Sci. Commun.* 7(1):1–9.

Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; Presser, S.; and Leahy, C. 2020. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Goes, F.; Zhou, Z.; Sawicki, P.; Grzes, M.; and Brown, D. G. 2022. Crowd Score: A method for the evaluation of jokes using large language model ai voters as judges. *arXiv preprint arXiv:2212.11214*.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4:246–279.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Premack, D., and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behav. Brain Sci.* 1(4):515–526.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds & Machines* 17:76–99.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1):92–96.

Sawicki, P.; Grzes, M.; Jordanous, A.; Brown, D.; and Peeperkorn, M. 2022. Training GPT-2 to represent two romantic-era authors: Challenges, evaluations and pitfalls. In *Proceedings of 13th International Conference on Computational Creativity*.

Searle, J. R. 1980. Minds, brains, and programs. *Behav. Brain Sci.* 3(3):417–424.

Stevenson, C.; Smal, I.; Baas, M.; Grasman, R.; and van der Maas, H. 2022. Putting GPT-3's creativity to the (alternative uses) test. In *Proceedings of the 13th International Conference on Computational Creativity*, 164–168.

Sturm, B. L. 2014. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia* 16(6):1636–1644.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ventura, D. 2016. Mere generation: Essential barometer or dated concept. In *Proceedings of the 7th International Conference on Computational Creativity*, 17–24.