Evaluating Prompt Engineering as a Creative Practice

Charlotte Bird
School of Design Informatics
University of Edinburgh
charlotte.bird@ed.ac.uk

Abstract

This short paper offers an overview of computational creativity evaluation methodologies that can be employed for the evaluation of prompt engineering. This task hopes to spark conversation around the role of computational creativity research in the new world of generative deep learning, and vice versa.

Introduction

The integration of new technologies into artistic practice is not a new phenomenon. The 1960s ushered in computers as an artistic medium, with institutions like The Museum of Modern Art and the Institute of Contemporary Arts[1] legitimizing the status of technologically entangled art.

Recently, multiple developments in deep generative modelling (Goodfellow et al. 2014; Ramesh et al. 2021; Ho, Jain, and Abbeel 2020; Rombach et al. 2022), have enabled new forms of human-computer creative interaction. The development of robust, consistent and adaptable generative models are powerful tools for creating new content such as text, images, games and music. Models like ChatGPT, DALL-E and StableDiffusion allow human-computer interaction and collaboration with little barrier to entry. Importantly, many users have employed these tools in creative processes.

Interaction with text-guided generative models is done through prompt engineering, or prompting. Prompting is the iterative development of textual commands which are designed to yield specific results. In the context of image generation, prompting has evolved into a creative process itself, and users can rapidly create impressive images. The accessibility and usability of text-to-image (TTI) models has precipitated the growth of hobbyists communities, adoption by professional artists and the creation of many peripheral resources. The popular communities and resources surrounding TTI systems are largely focused on refining prompting practice through sharing or buying prompts, sharing trained models and outputs, and offering advice on developing prompting processes.

The field of computational creativity has, for a long time, been discussing the questions, insights and problems that arise from creative interaction with computers. However, the generative deep learning field has yet to implement such findings with a view to evolving generation systems. Evaluation is a primary example of this. The CC field has well-developed evaluation methodologies designed to capture instances of creativity, improve systems and identify progress, yet not one has been utilised, or even the connection made.

This short paper intends to build upon these initial findings to apply evaluation frameworks originally developed to identify how and where systems exhibit "computational creativity". In doing so, this paper is designed to ignite a conversation about what generative deep learning can learn from computational creativity, and what computational creativity can learn from the development and mass use of systems that are "creative" but not explicitly *computationally creative*. This is achieved through the application of evaluation frameworks originally developed to identify how and where systems exhibit "computational creativity".

It is important to note that the computational creativity (CC) field has already extensively discussed the questions, insights and problems that arise from creative interaction with computers. However, the generative deep learning field has yet to implement such findings with a view to evolving generation systems. Evaluation is a primary example of this. The CC field has well-developed evaluation methodologies designed to capture instances of creativity, improve systems and identify progress, yet not one has been utilised, or even the connection made.

Related Work

Margaret Boden initially proposed novelty and value as desired criteria in computational creativity tasks (Boden 1998; 2004). Ritchie subsequently proposed a summative evaluation method by judging the product of creative systems for typicality/novelty and quality (Ritchie 2007). Colton (Colton 2008) alternatively emphasises the importance of process through assessing the presence of three criteria: *skill, imagination* and *appreciation*. Later, the FACE/IDEA models were designed to describe and capture the impact of creative acts (Colton, Charnley, and Pease 2011). The SPECS system (Jordanous 2012) was developed to resolve the need for clear and defined benchmarks across com-

[1] The Machine as Seen at the End of the Mechanical Age, The Museum of Modern Art, New York, 1968-1969, Cybernetic Serendipity, Institute of Contemporary Arts, London, 1968.

putational creativity evaluation. SPECS evaluates systems against 14 factors identified through creativity studies. Evaluating computational creativity systems can also be undertaken via Turing-style comparison tests (Pearce and Wiggins 2001; Boden 2010), though such tests are criticised (Pease and Colton 2011).

Prompt engineering research is limited. Current work includes a six-type prompt taxonomy (Oppenlaender 2022b) and prompting design guidelines (Liu and Chilton 2022). Additionally (Oppenlaender et al. 2023) investigate perceptions of TTI generation, such as possible applications, dangers and concerns. A number of authors have explored the *skill* of prompt writing (Chang et al. 2023; Oppenlaender, Linder, and Silvennoinen 2023; McCormack et al. 2023).

Evaluating Prompt Engineering

This section offers the beginnings of a discussion for the evaluation of prompting. This discussion is in light of recent prompting research that bypasses meaningful evaluation (Chang et al. 2023), even if such evaluation is borrowed from the CC field or otherwise.

Product

Image The goals of text-to-image systems such as DALL-E is to generate images according to a given prompt. An essential sub-goal is the generation of images that properly express the creative and aesthetic aims the users expresses via the prompt. The user will likely seek to generate subjectively novel and quality images, though the achievement of this goal is contentious, which I later discuss. Despite this, the image is still an interesting object of discussion. In online communities, users share their images according to themes (sci-fi, fantasy, horror, photography, etc), where they can receive feedback or praise.

Prompt A secondary aim within TTI communities is the creation of novel and valuable *prompts*. This sub-goal is achieved sometimes, and is validated through the sharing and sale of prompts [2]. Significant value is often ascribed to the prompt as part of the "artwork" (Chang et al. 2023), however novelty and value in the prompt is entirely distinct from a novel and valuable image: though the two are commonly conflated. As with the image output, the legitimacy of prompt engineering as a skill is contended (McCormack et al. 2023), though some argue that experience with the training set, the models latent space and using particular prompt modifiers evidences a skill (Oppenlaender 2022a). Creating novel and valuable prompts relies on a novel approach to linguistic expression and traversing the latent space. An artist who is able to express a vision through the use of unexpected and surprising prompts evidences more skill than a user who is able to cycle through prompt modifiers, even if the latter produces "better" images.

Portfolio The ability to rapidly generate and edit high-quality images allows users to quickly build portfolios. Where it may take an artist 5 years to develop a sizeable body of work, a user could dedicate a day. The curation of an aesthetic and style within a portfolio is another way a user may exert creative control. Prominent "AI artists" cultivate a specific style, which they often mint as NFTs and try to sell.

Evaluation Boden (Boden 2004) makes the important distinction between P-creativity (novel to creator) and H-creativity (novel to culture). In the context of prompting and generation, we concerned with the production of novelty relative to its initial state of knowledge (P-creative) (Ritchie 2007). and we can relate the P-creative to the individual and community generating the prompts. Ritchie's development of 14 (later 18) criteria defines three key mappings: *novelty* and *typicality* in the intended domain and *value* of the output (Ritchie 2007; Boden 1998). Ritchie defines an inspiring set I, wherein the formal account of creativity is judged according the replication or imitation of I. Suitably novel outputs V are therefore derived from the output set O. The degree of creativity is determined by the number of novel output V produced which are not in I (Colton et al. 2002). "Fine-tuning" (Colton et al. 2002) is when systems evidence replication to a greater extent than the generation of novel high-value items. It has been proven that systems such as Stable Diffusion generate statistical amalgamations of the dataset, evidencing reconstructive memorization and imitation (Somepalli et al. 2022). This is not always easily recognised due to sheer size of the datasets. The prompt as output also evidences such limitations, many prompts that utilise guides or common modifiers are fundamentally not novel, and as such novelty and value arises in the unexpected use of language, which is arguably finite and bound by linguistic limitations (McCormack et al. 2023). Qualitatively, we can argue for novelty in output (i.e this image has not existed before), however quantitatively, it is proven that true novelty (not the imitation of) in contained TTI generation is difficult given the limitations of only rendering that which always exists (McCormack et al. 2023). However, we are also able to consider the presence of value in the form of writing the prompt if we consider the prompt as a novel creative act. We could consider the prompt process as akin to writing a series of exploratory questions.

To argue either side is to decide whether such forms of creation predicated on amalgamation, imitations, pastiche and mimicry (even possibly *unrecognisably* so) can ever represent novelty. Importantly, this does not hold for individual creative processes, only simplified prompt engineering. This exemplifies the current divide in TTI research (Chang et al. 2023; McCormack et al. 2023).

Clearly, Ritchie's criteria present a number of theoretical issues: such aesthetic measures are highly subjective and practically difficult to implement, and offer no answers for evolving generated outputs to evidence novelty without expanding the capabilities of the system beyond the inspiring set. With value and novelty contentious criteria, the IDEA ((I)terative (D)evelopment(E)xecution-(A)ppreciation) de-

[2] promptbase.com

scriptive model (Colton, Charnley, and Pease 2011) offers a second path of product evaluation. The IDEA model is composed of two tasks. The first describes the stage of development, the second posits the *impact* of creation as opposed to the value metric. The IDEA model supposes an ideal audience *(i)* and quantitatively measures the impact a creative act *(A)* has on *i*. Disregarding the subjective metric simplifies many of the problems attached with evaluating prompt and image. Instead, we evaluate according to the ideal audience. In evaluation of the prompt and image as a mutually reinforcing art piece, we ideally evaluate the outputs (prompt and image) according to their proximity to each other and the dataset. Ideally this measure includes (for example) shock and subversion. Additionally, the IDEA model supposes two further simplifying solutions: ideal development process and ideal background knowledge information, which, alongside creating the ideal audience, may be as challenging as generating the creative artefact.

The application of evaluative methodologies to prompt engineering is messy at best. The above examples have aimed to show just how difficult it can be to define exactly how we think about value and novelty as requirements of creativity, especially in closed generative models. In addition, a user may rate their outputs as novel, valuable, unexpected or appealing, and therefore call themselves an artist. Indeed many in the community do. Therefore any possible evaluation of product must not *rely* on the user self-assessing, as has been done in previous studies (Chang et al. 2023), but must consider evaluation by expert users and audience. The prevalence of self-assessment and validation has only supported the criticisms levied at the communities, such as in artistic theft. However, preliminary analysis of online communities reveals a growing body of users who consider their outcomes valuable and original. This conclusion is largely premised on their reluctance to employ existing style words, artist names and over-used prompt techniques. As such, they would be an interesting place to start with this evaluation.

Process

The creative process can be broken down into a number of stages, for example preparation, incubation, illumination, and verification (Wallas 1926). Prompt engineering processes do not fundamentally differ from other creative processes, except that some stages (or tasks) are undertaken by a generative model. Much of the process is also undertaken as an iterative interaction between human user and model.

Iteration Prompt engineering has previously been broken down into two native tasks: iteration and curation. The central goal of the iteration task is to refine textual descriptions according to the previous generation in order to reach a desired image. Users must navigate and map the model's latent space via text, often times finding strange, seemingly unrelated connections or glitches. The iteration process is co-creative as both user and machine contribute to the problem solution, and should be evaluated as such. (Chang et al. 2023) found that a common goal for users was to pursue new capabilities through the creation of a specific visual language: employing words from differing domains alongside their natural vocabulary. The user's creative logic and expression is altered by the billions of text-image mappings.

Curation Users may curate an image or images through editing techniques such as inpainting, outpainting or retouching. Image synthesis models frequently fail to properly render spatial arrangements, faces or text, or simply do not achieve the goals of the prompt. Often, the creative aims of the image are reached within iteration, and curation simply resolves the expected failings of the generator. However, curation can also be a creative task. Artists may use the curation phase to exert creative agency through minimal or extensive editing, such as involving other mediums and tools, or using the generated image as inspiration or a fragment of a larger creative vision. One artist uses generative models to produce human forms, which are then painted over, another uses them to create portions of a collage[3]. The exertion of creative agency by the user is a oftentimes where value arises. Framing information - such as intention or process - are key to legitimising the final image as the result of a meaningful creative act, rather than mere generation.

Collaboration It is difficult to quantify the influence of the community on the artist. From a distance, it is possible to see how new techniques, styles and subjects disseminate, however in proximity, art appears a nebulous and interwoven world. Prompt engineering offers an unusual insight into collaborative creativity. We can call this an instance of P-creativity In the many Discord-based communities, users directly copy prompts, images, techniques, applications and ideas (Oppenlaender 2022a). This collaboration is uniquely supported by the inability to copyright AI-generated images, and the extremely low barrier to entry: anyone write prompts and contribute to the community. It is important to recognise that artists whose style, name and artworks are adopted by the prompting communities are also - unwillingly - drawn into this process.

Evaluation When we appreciate an artefact, we are appreciating both the process and the outcome (Colton 2008). We acknowledge the skill, time, dedication, knowledge and application of the artist. Our perception of how something is produced can influence our reception of the outcome (Colton 2008). This is especially applicable to prompt engineering, wherein we may misjudge or undervalue an artefact because we believe it is generated. Traditional artists have taken to posting their process - framing information - to prove their work is not generated. The digital image and prompt tell us little about the skill of the creative process and therefore user autonomy in the creative process is highly valuable. A user may have copied a prompt, utilised prompt "cheat words", or simply have stumbled upon a good output. At the same time, the user may have undergone an extensive iteration and

[3] These insights were gained from personal conversations with artists.

curation process, guided by a focused creative vision. Similarly, users can fork, train and alter TTI systems to their own specifications, which can be a creative skill in itself. Importantly, it is not fruitful nor useful to apply process evaluation to the act of simply typing a prompt - "bear in a suit" - and generating an image. We expect a user to employ and defend some creative process, artistic, linguistic, collaborative or otherwise.

To evaluate prompting processes, we must employ multiple evaluative methodologies. The first is utilised for the evaluation of mixed-initiative co-creativity, and aims to quantify the degree of use of the generated images and the quality of use within the path of creation (Yannakakis, Liapis, and Alexopoulos 2014). In this case, the goal of assessing the degree and quality of use is to conclude whether or not the generative model fosters or undermines the creativity of the user. We ideally want to understand the quality of use (understanding, subversion, evolution, exploration, re-appropriation) through asking the subsequent questions. For example, a human audience can be used to reveal the usefulness of TTI systems outside of mere generation: how useful are they in iterating through ideas? Can we identify milestones (Yannakakis, Liapis, and Alexopoulos 2014) where the user feels creatively undermined or supported? It is also important to quantify the impact that communal co-creation has on an individual user. For example, how are creative processes undermined or enhanced by community resources? How does a user seek and identify novelty and value in their product in light of the limitations of textual commands (Chang et al. 2023).

The FACE model (Colton, Charnley, and Pease 2011) captures and emphasize the importance of the process of artefact creation in a judgement of creativity. Prompt engineering and generation express multiple instances of individual generative acts. The TTI system performs creative acts of the form C^g, E^g as an executable program and subsequent execution by the user. It is also possible to argue that process evidences A^g (a local aesthetic) as images and prompts are judged according to a users given heuristics. F^g or *framing information* (natural language text that is comprehensible by people) in the form of the prompt or user-provided explanations arguably adds value by imbuing the output with meaning, and linking them to some human motivations The FACE model can be used comparatively, for example $<C^p> \; > \; <A^g> \; > \; <C^g> \; > \; <E^g>$, wherein we prioritise the method of generating concepts, in which prompting and generation may score badly, as we can argue that the generation act is imitation. It is also suggested to utilise the FACE model $CA1 = <C^g, E^g> \; < \; CA2 = <A^g, C^g, E^g>$ where the invention or choice of an aesthetic by the computer is 'more creative'. Unsurprisingly, it is difficult to apply the FACE model to a process we have categorised as co-creative. However we can apply aspects, such as prioritising the method of generation and the invention of an aesthetic in evaluating process. It would be interesting to evolve the TTI process by enabling greater agency through the generation of framing information, for example.

The final evaluative methodology is the Creative Tripod (Colton 2008). Colton argues that in order for software to be perceived as creative, it should display three behaviours: *skillful*, *appreciative*, *imaginative*. Whilst a more simplistic approach, this framework can provide insights into developing prompt engineering and generation. Any party can also contribute to the tripod (programmer, consumer and computer). If we extend this to consider the user, we can argue that a user frequently evidences skilful interaction with the system through prompts, though the skill may not yield creative or valuable results. Whilst we cannot call the TTI system "appreciative" or "imaginative", we can recognise the fine-tuned capabilities of the system to generate impressive, realistic and artistic images. An inclusion of Ritchie's criteria (Ritchie 2007) to evolving such capabilities may also yield more "creative" processes. Application of the tripod to prompt engineering is difficult, as we largely care about the co-creative relationship rather than the empty appearance of creativity.

Problems, Criticisms and Future Work

Artistic endeavour frequently manifests as divergence away from established mediums, forms, tools, techniques and subjects. Many cite the *Portrait of Edmond Belamy* and the brief popularity of NFTs as watershed moments in how artists can create art, and how customers purchase it. However this acceptance of a new suite of technologies has ignored many legal and ethical concerns. In addition, "AI art" is not always well received. Job displacement, market saturation, data laundering, copyright infringement and artistic legitimacy are only some of the issues up for debate. In addition to such concerns, it is often argued that the act of prompting does not diverge from the previous method of human-computer interaction via textual commands, and is limited to the combination or exploration of defined concepts or objects which can be expressed via natural language (McCormack et al. 2023). In this way, prompt engineering is akin to a database query. Further, it is easy to buy prompts or even generate them[4]. Further, TTI systems are dependent - even *parasitic* (McCormack et al. 2023) - on existing and new human visual data to generate 'new' images, without which outputs would devolve into pastiche.

Considering this, it is difficult to foresee widespread acceptance of prompt engineering as an *artistic* practice. Yet, it is likely that the adoption of such tools will only increase. It is important to note that the combination and expression of concepts via natural language is a foundation of human knowledge production, and undermining prompt engineering as a creative practice because of the limitations of language would undermine countless creative acts. By extension, prompt engineers are well-supported in calling their process creative yet it is interesting to consider the forms of divergence that could legitimise the process. For example, where a system is altered to provide debate or increased interactive tangibility[5] rather than mere generation. Divergence could also be realised when a user subverts the in-

[4] https://huggingface.co/succinctly/text2image-prompt-generator

[5] One artist mentioned that without two-way digital or physical interaction the process does not feel creative.

tended use of the model, exposes or alters the fundamental processes of generation.

Conclusion

This paper is a preliminary discussion of what generative deep learning can learn from a CC perspective, from the view of evaluation. I would suggest that generative deep learning has largely ignored the CC literature in system development because they do not consider creativity as a compelling aspect of generation interactions, rather focusing on developing "better" systems. I have hoped to show that CC evaluation offers a method to assessing system limitations, whilst also offering insights as to developing systems to better assist in (co-)creativity. For example, this paper has presented a number of failings in generative models such as pastiche and imitation, limited interaction and opaque process. It is also important that the CC field considers what it can learn from the mass use of deep generative models in a creative context, as these new interactions offer ripe opportunity to understanding the processes and interactions of the user with "creative" systems. This work is presented with the intention to pursue further analysis, but I have hoped to exemplify some of the connections to be made between the fields.

Acknowledgments

I would like to thank the anonymous reviewers for their kind and constructive feedback. This research was supported by AHRC BRAID programme AH/X007146/1 and the Baillie Gifford PhD scholarship at the Centre for Technomoral Futures.

References

Boden, M. A. 1998. Creativity and artificial intelligence. *Artificial Intelligence* 103(1):347–356.

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. Psychology Press.

Boden, M. A. 2010. The Turing test and artistic creativity. *Kybernetes* 39(3):409–413.

Chang, M.; Druga, S.; Fiannaca, A.; Vergani, P.; Kulkarni, C.; Cai, C.; and Terry, M. 2023. The Prompt Artists. arXiv:2303.12253 [cs].

Colton, S.; Pease, A.; Ritchie, G.; and Bridge, S. 2002. The Effect of Input Knowledge on Creativity. *Technical Reports of the Navy Center for Applied Research in Artificial Intelligence.*

Colton, S.; Charnley, J.; and Pease, A. 2011. Computational Creativity Theory: The FACE and IDEA Descriptive Models. In *Proceedings of the 2nd International Conference on Computational Creativity.*

Colton, S. 2008. Creativity Versus the Perception of Creativity in Computational Systems. In *AAAI Spring Symposium: Creative Intelligent Systems.*

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. arXiv:1406.2661.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

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. Place: Germany Publisher: Springer.

Liu, V., and Chilton, L. B. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *CHI Conference on Human Factors in Computing Systems*, 1–23. New Orleans LA USA: ACM.

McCormack, J.; Cruz Gambardella, C.; Rajcic, N.; Krol, S. J.; Llano, M. T.; and Yang, M. 2023. Is Writing Prompts Really Making Art? In *Artificial Intelligence in Music, Sound, Art and Design: 12th International Conference, EvoMUSART 2023*, 196–211. Springer-Verlag.

Oppenlaender, J.; Visuri, A.; Paananen, V.; Linder, R.; and Silvennoinen, J. 2023. Text-to-Image Generation: Perceptions and Realities. In *Workshop on Generative AI in HCI (CHI '23')*.

Oppenlaender, J.; Linder, R.; and Silvennoinen, J. 2023. Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering. arXiv:2303.13534.

Oppenlaender, J. 2022a. The Creativity of Text-to-Image Generation. In *25th International Academic Mindtrek conference*, 192–202. arXiv:2206.02904.

Oppenlaender, J. 2022b. A Taxonomy of Prompt Modifiers for Text-To-Image Generation. arXiv:2204.13988.

Pearce, M., and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. *Proceedings of Symposium on Artificial Intelligence and Creativity in the Arts and Sciences.*

Pease, A., and Colton, S. 2011. On Impact and Evaluation in Computational Creativity: A Discussion of the Turing Test and an Alternative Proposal. In *AISB 2011: Computing and Philosophy.*

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. arXiv:2102.12092 [cs].

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. *Minds and Machines* 17(1):67–99.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752 [cs].

Somepalli, G.; Singla, V.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2022. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. arXiv:2212.03860 [cs].

Wallas, G. 1926. *The Art of Thought*. Harcourt, Brace.

Yannakakis, G. N.; Liapis, A.; and Alexopoulos, C. 2014. Mixed-initiative co-creativity. In *Proceedings of the 9th conference on the foundations of digital games.*