A Small-Data Mindset for Generative AI Creative Work

GABRIEL VIGLIENSONI, Creative Computing Institute, University of the Arts London, UK
PHOENIX PERRY, Creative Computing Institute, University of the Arts London, UK
REBECCA FIEBRINK, Creative Computing Institute, University of the Arts London, UK

In this paper, we argue that working with small-scale datasets is an often-overlooked but powerful mechanism for enabling greater human influence over generative AI (GenAI) systems in creative contexts. We describe some of the benefits of working with small-scale data, and we argue that conventional ways of thinking about the value of large data, such as preventing overfitting, are not always well-matched to creative aims. We discuss how models built with small-scale data can facilitate meaningful creative work, providing examples from text, image, and sound.

CCS Concepts: • **Computing methodologies** → **Neural networks**; • **Applied computing** → **Arts and humanities**; • **Human-centered computing** → **Human computer interaction (HCI)**.

Additional Key Words and Phrases: small-data, generative systems, creative AI

ACM Reference Format:
Gabriel Vigliensoni, Phoenix Perry, and Rebecca Fiebrink. 2022. A Small-Data Mindset for Generative AI Creative Work. In *Generative AI and HCI - CHI 2022 Workshop, May 10, 2022, Online* . ACM, New York, NY, USA, 5 pages. https://doi.org/10.5281/zenodo.7086327

1 INTRODUCTION

Generative AI (GenAI) systems capable of outputting novel text or media content have captured the interests of both creative artists and the public, especially following the past decade's advancements in autoencoders, GANs, and transformers. Artists and designers employing such systems may release trained generative models into the world, with their unfiltered outputs fully on display (e.g., [6]); they may heavily curate model outputs to choose which to release as creative pieces or components thereof (e.g., [17]); they may interactively manipulate models to performatively traverse their output spaces (e.g., [1, 19]); or they may do something else entirely.

Many popular approaches to generation of text or media employ very large neural network architectures and require massive datasets to be trained. However, in creative applications, this is sometimes in conflict with a human creator's desire to generate content which is unusual, personalised, or otherwise distinctive, and/or with creators' lack of access to large datasets and massive computing power. In this position paper, we argue that the nearly ubiquitous assumption that bigger data is better in generative AI systems is not always appropriate, particularly in creative use contexts. We point out some of the ways that creators' goals and ways of working can be better served by "small data", drawing on examples of creative practice in various domains.

Manuscript submitted to ACM

2 WHY SMALL-SCALE DATA?

2.1 Creative ML may have different goals

Commonly, ML is understood as a tool for building models that must satisfy a large number of people or work in a large variety of contexts, and therefore benefit from a large amount of data to be trained. In reality, though, creative AI systems are often intended to be more niche and may be customised to the needs of one maker, and potentially just for one specific purpose. For example, in music, Dadabots [4, 22] have adapted underground music's countercultural tradition of appropriating and remixing source material. They have used SampleRNNs [13] trained on a single existing album to generate music "within the limited aesthetic space of the album's sound" [4]. As another example, Anna Ridler's "Fall of the House of Usher" [15] used a small dataset of Ridler's own drawings to train a pix2pix model to translate frames from a film into new "drawings" in her style.

Additionally, as described above, many creators employ GenAI systems within larger human creative contexts, e.g., curating outputs or performatively interacting with models in ways that draw on their strengths. This means that in some contexts, it may not be very important that all (or even most) outputs of a generative model are "realistic", or that a model is capable of generating outputs with a huge range of variety, or that the latent space of a model is good at capturing certain structural features. One consequence is that, in some creative applications, makers may not benefit as much from properties of large models trained on large data that make these attractive in other domains.

2.2 Small data can mean less—and more effective—human work

In creative usage contexts, one can understand the choice of training dataset as the primary mechanism by which a human creator specifies what kind of content the machine should generate—e.g., should it generate black metal music, or Beatles' "number one" hits [4]? Should it generate images in the style of the human artist's own work [15] or shall it generate photorealistic seascapes [1]? When this specification requires only a small number of examples, this translates into less work for creators who must generate or curate examples.

When datasets are small, small changes such as the addition or removal of a few training examples can also have appreciable effects on the trained model's behaviour. Further, simpler architectures capable of being trained on small datasets can often be trained faster. This means that not only is less computing power and expense required, but creators can also more easily iterate—e.g., repeatedly changing training examples in order to interactively steer a model toward a desired behaviour. This is key to enabling people to exercise creative agency during model creation. For example, [20] scaled-down a large VAE architecture to be able to train a generative model of music rhythms faster than previous attempts. Instead of learning one big, complex model, they took a small-scale approach to learn several smaller, more manageable models that allowed them to iteratively explore and improve each one.

A related approach to taking advantage of smaller datasets which can also help creators "steer" the behaviour of GenAI systems through the choice of data is the fine-tuning of large, pre-trained models. Instead of retraining large architectures from scratch, the pre-trained weights of the original dataset are used as a starting point when new data is processed in further training iterations. For instance, the "ReRites" project by Johnston [12] involved fine-tuning the large pre-trained language model GPT-2 on his custom contemporary poetry corpus to generate poems in realtime and present them as an installation,[1] as well as to make a series of twelve books in which Johnston edited the model's output [12]. In music, the artists Hexorcismos and Dadabots fine-tuned the large Jukebox transformer model to create

[1] https://vimeo.com/335698694?embedded=true&source=video_title&owner=4131166

24/7 AI raves[2] using EDM sets from Dublab Radio, and to reimagine 3ball music, a Mexican traditional style, based on a 3ball dataset collected from YouTube as the training material [11]. While fine-tuning is a key enabler of some such work, we emphasize that it is not always the right tool. Some creative work alternatively benefits from data augmentation techniques applied to smaller datasets (e.g., [1]) and other work with small data (e.g., [20]) requires no such measures.

2.3 What about bias?

In one sense, the "bias" inherent in a small, manually-curated, well-understood dataset may be desirable, when a creator is employing this dataset as a mechanism for communicating to the machine exactly what sorts of outputs she would like to generate.

Further, requiring large datasets introduces its own form of bias, in that we are limited to modelling phenomena for which large amounts of data can easily be sourced. For example, very large language models such as GPT-3 [3] and SWITCH-C [9] have been trained on large, not-properly curated, static data dumps from the Web. Bender et al. argue that this indiscriminate approach to data collection poses risks because it encodes hegemonic worldviews, amplifies biases and other problems in the training data, and can be harmful to marginalized populations when deployed and used at large [2]. The large data required to train massive models may introduce normative biases in other disciplines as well. For example, the generative model for music Jukebox [7] was trained on more than a million songs, half of which were in English, and with a heavy bias toward mainstream music genres, making it difficult (in our experience) to use it as it is with more niche genres. As another example, large data endeavours also struggle to perform movement modelling for disabled bodies due to the uniqueness of disabled people and the tendency of developers to design for the majority and ignore outliers [21].

GenAI art can in fact be an interesting vehicle for revealing and critically engaging with bias as a phenomenon in machine learning. For example, Ridler writes about how seeing the GAN-generated images in "Fall of the House of Usher" drew her attention to hitherto-unnoticed patterns in her own drawings [15]. Memo Akten's "Learning to see"—a pix2pix-based piece in which a neural network must translate what it "sees" in a human-controlled camera feed (e.g., a set of keys) into the one type of image it has been trained to produce (e.g., seascapes)—is motivated in part by a desire to expose biases and shortcomings in neural networks [1].

2.4 What about overfitting?

As described above, often the goal of a GenAI system is not to create a model capable of making something wholly "new" or "creative" in the way a human would, but to remix, recast, or remake within a more limited scope. For instance, Dadabots describe desiring to "overfit short timescale patterns (timbres, instruments, singers, percussion) and underfit long timescale patterns (rhythms, riffs, sections, transitions, compositions) so that it sounds like a recording of the original musicians playing new musical compositions in their style" [4]. In the visual domain, overfitting may be also desired when trying to have a generative system producing something recognizable, as in the case of "Mosaic Virus" by Ridler where a GAN is trained on a bespoke dataset of tulips to generate artificial ones on screen whose characteristics mutate and morph depending on unseen financial markets' fluctuations [16].

[2] https://aimusicfestival.eu/en/programs/2021/areas/concerts-amp-demo-research/ai-rave-by-dadabots-x-hexorcismos-dataset-by-dublab

2.5 Reclaiming power

Building one's own small-scale datasets for creative ML can be seen as one way of subverting the power structures that reign AI development. First, we bypass the large corporations and institutions collecting and building datasets—some with ethically questionable underpinnings—who are also in control of the processing power needed to train large models. Second, there is arguably a history of tech companies enticing digital artists to learn, master, and explore the design space of new technologies but then locking down or charging significantly more for access to those technologies once artists have served the purpose of promoting those technologies to the wider public (e.g., see the history of Microsoft's Kinect in digital art). Third, we take a diversion from the "artistic canon" that validates and perpetuates the specific examples or aspects of (usually one) culture that are established as paradigmatic and crucial.

3 RELATIONSHIP TO OTHER WORK

Research on creative uses of supervised learning by Fiebrink et al. has similarly pointed out how the use of small datasets can be useful in enabling "interactive machine learning" [8] in creative contexts, for instance where a gesturally-controlled musical instrument may be built from paired examples of gestures and sounds. In such contexts, small data can enable model "steering" toward subjective goals (e.g., creating interfaces that are expressive and comfortable), and this also changes considerations about ML infrastructure (e.g., algorithms that "overfit" such as kNN can be preferable in that they enable learning of complex functions from smaller datasets).

On the other hand, significant current research in GenAI focuses on very large architectures that maintain the same underlying model for different tasks (e.g, OpenAI's Codex [5] generates code, and Dall·e [14] generates images, but both use GPT-3 as their foundation). As such models grow, their need for more training data increases along with problems such as bias and environmental costs, as researchers such as [2] have warned. Data-Centric AI[3] is a research initiative that advocates for prioritizing data quality instead of quantity, arguing that focusing on high-quality data—by engineering and working with small subsets of the data—can address model biases in a more targeted way [18].

4 CONCLUSION

In this paper, we have argued for the value of a "small-data" mindset for generative AI in creative contexts. We do not claim that "big data" is always bad, or that generalisation (or the generation of totally new and yet coherent content) is never important. However, we do argue that assumptions about the value of big data and big models are worth questioning, lest we overlook other approaches to making GenAI useful in creative practice. Small data and models bring their own challenges, such as the lack of ability to generate widely varying outputs, but they also enable makers to play with technology more directly, leaving space to reflect on topics about the relation between art and technology, co-creation with machines, the role of the creator, errors in modelling reality, and creative workflows among others.

Whereas state-of-the-art AI algorithms are often open-source and well documented, the datasets used to train large models—and the methodologies to collect and prepare them—are often not readily available. And even if they are available, their sheer size makes it virtually impossible to revise every single data point or example. In such a context, creative practitioners miss out on the ability to exert influence over the reality a model is trying to mimic: the data. As a result, they can only experiment and play with a given model of a (problematically biased) depiction of reality. A small-data mindset for GenAI systems, where we recognise the value of smaller models trained on smaller but well-curated datasets, can help to change that.

[3] https://datacentricai.org/neurips21/

REFERENCES

[1] Memo Akten, Rebecca Fiebrink, and Mick Grierson. 2019. Learning to see: You are what you see. In *ACM SIGGRAPH 2019 Art Gallery*. ACM, Los Angeles, CA, 1–6. https://doi.org/10.1145/3306211.3320143

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, Montréal, QC. Canada., 610–623. https://doi.org/10.1145/3442188.3445922

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. https://doi.org/arXivpreprintarXiv:2005

[4] C. J. Carr and Zack Zukowski. 2018. Generating albums with SampleRNN to imitate metal, rock, and punk bands. In *Proceedings of the 6th International Workshop on Musical Metacreation (MUME 2018)*. http://arxiv.org/abs/1811.06633

[5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. (2021). http://arxiv.org/abs/2107.03374

[6] Dadabots. 2019. Relentless doppelganger. https://www.youtube.com/watch?v=MwtVkPKx3RA

[7] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341* (2020).

[8] Jerry Alan Fails and Dan R. Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*. Miami, FL, 39–45.

[9] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. (2021). http://arxiv.org/abs/2101.03961

[10] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 147–156.

[11] Moisés Horta-Valenzuela. 2021. Okachihuali, by Hexorcismos. https://hexorcismos.bandcamp.com/album/--2

[12] David (Jhave) Johnston. 2019. *ReRites*. Anteism Books, Montréal, QC. https://www.anteism.com/shop/rerites-david-jhave-johnston

[13] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2017. SampleRNN: An unconditional end-to-end neural audio generation model. In *International Conference on Learning Representations (ICLR 2017)*. http://arxiv.org/abs/1612.07837

[14] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. (2021). http://arxiv.org/abs/2102.12092

[15] Anna Ridler. 2018. Fall of the House of Usher. Datasets and decay. https://www.vam.ac.uk/blog/museum-life/guest-blog-post-fall-of-the-house-of-usher-datasets-and-decay

[16] Anna Ridler. 2019. Mosaic Virus. http://annaridler.com/mosaic-virus

[17] Tom Simonite. 2017. A 'Neurographer' Puts the Art in Artificial Intelligence. https://www.wired.com/story/neurographer-puts-the-art-in-artificial-intelligence/

[18] Eliza Strickland. 2022. Andrew Ng: Unbiggen AI. https://spectrum.ieee.org/andrew-ng-data-centric-ai

[19] Gabriel Vigliensoni. 2021. Clastic music, MUTEK 2021. https://mutek.org/en/artists/vigliensoni-1

[20] Gabriel Vigliensoni, Louis McCallum, and Rebecca Fiebrink. 2020. Creating latent spaces for modern music genre rhythms using minimal training data. In *Proceedings of the 11th International Conference on Computational Creativity (ICCC'20)*. Coimbra, Portugal (Online).

[21] Mike Wald. 2021. AI Data-Driven Personalisation and Disability Inclusion. *Frontiers in Artificial Intelligence* 3 (2021). https://www.frontiersin.org/article/10.3389/frai.2020.571955

[22] Zack Zukowski and C. J. Carr. 2017. Generating Black Metal and Math Rock: Beyond Bach, Beethoven, and Beatles. http://arxiv.org/abs/1811.06639