

Large-scale generative simulation artificial intelligence: The next hotspot

Qi Wang,¹ Yanghe Feng,³ Jincai Huang,^{3,*} Yiqin Lv,² Zheng Xie,² and Xiaoshan Gao^{4,*}

¹Kaiyuan Mathematical Sciences Institute, Changsha 410000, China

²College of Science, National University of Defense Technology, Changsha 410073, China

³College of Systems Engineering, National University of Defense Technology, Changsha 410073, China

⁴Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

*Correspondence: huangjincai@nudt.edu.cn (J.H.); xgao@mmrc.iss.ac.cn (X.G.)

Received: July 22, 2023; Accepted: September 13, 2023; Published Online: September 19, 2023; <https://doi.org/10.1016/j.xinn.2023.100516>

© 2023 This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Citation: Wang Q., Feng Y., Huang J., et al., (2023). Large-scale generative simulation artificial intelligence: The next hotspot. The Innovation 4(6), 100516.

MOTIVATION

Nowadays, big data, deep-learning models, optimization methods, and computational power are essential elements in promoting the development of artificial intelligence. Recent advances have brought a new focus on generative artificial intelligence (GenAI), which paves promising paths to exploring the creation of texts, images, videos, or other content rather than simply performing discriminative learning tasks.

GenAI's emergence, e.g., the large model, has changed the landscape of deep-learning research and has influenced individuals' work and life, holding tremendous potential to reshape robotics research, national governance, and life sciences. Consequently, a pressing question arises: "How will GenAI inspire technological revolution?" In answer to this question, this commentary highlights the fascinating functions of GenAI, stresses the importance of experimental design, discusses critical inductive biases as the geometric prior, identifies multi-views in the evaluation, and nominates large-scale generative simulation AI (LS-GenAI) as the next hotspot for GenAI to connect. Throughout the commentary, we use $x \in X$, $y \in Y$, and $z \in Z$ to denote the explanatory variable, the response variable, and

the latent variable, respectively. For tasks or operators on datasets τ , we represent them in distributions $p(\tau)$.

GenAI can do more than AIGC

The widespread popularity of GenAI models stems from their ability of AI-generated content (AIGC). Technically, the deep generative model empowers GenAI's numerous utilities beyond standard AIGC. Among them, we list three practical ones in Figure 1A.

Minimizing the number of required bits to store and transmit information, known as data compression, is crucial. This utility is particularly essential in time-sensitive services with memory constraints, such as edge computing. Some generative models, such as the vector-quantized variational autoencoder, excel in data compression by finding insufficient statistics of high-dimensional signals. Representation disentanglement refers to the ability to infer statistically independent latent variables that explain different aspects of data generation, e.g., style, color, and pose. It closely relates to the controllable generation, e.g., obtaining samples with only one aspect varied. Causality is also an arising consideration in GenAI, and generative models are advantageous in handling

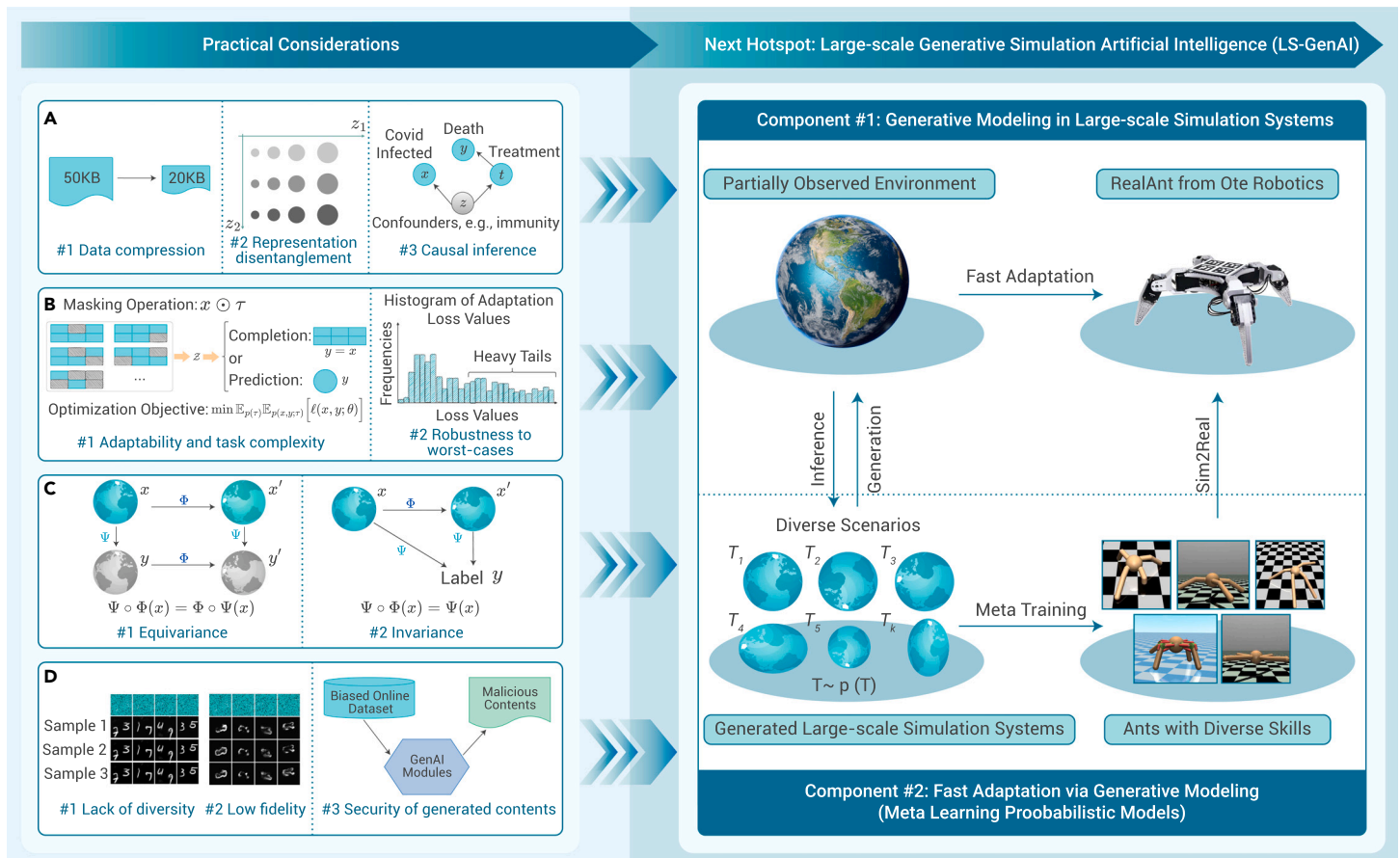


Figure 1. Practical considerations and the roadmap of LS-GenAI The left supports the utilities of LS-GenAI, while the right is the roadmap to achieve via the doubly generative paradigm.

high-dimensional variables, discovering causal graphs, and enabling counterfactual predictions. For example, $p(y|x, do(z = z_0))$ can render the influence of socio-economic policies denoted by z_0 without incurring additional costs.

Despite these fascinating utilities, there remain several tricky issues in the field, such as the plausibility of representation disentanglement and causal discoveries in the presence of small-scale datasets.

Experimental design matters in GenAI's adaptability and robustness

Let us rethink the critical factors contributing to GPT-like models' success. The language's generative process must be capable of capturing the masked input-output coupling pattern in the corpus, mapping linked entities to a knowledge earth, and continually updating with new patterns. Hence, when users initiate queries for specific contextual terms, the built knowledge earth can effectively locate and feedback the precise information.

The above process inspires the task distribution design for GenAI. Task diversity nurtures models' generalization capability across various scenarios in zero-shot or few-shot learning. However, increasing task complexity requires larger model sizes and consumes more computations. For instance, GPT-3 has 175 billion parameters and has been trained on over 570 GB text data from numerous tasks. Masked learning has emerged as one of the most popular heuristics in task generation. Nevertheless, exhaustively exploring all scenarios in training large models can be computationally demanding. As an example, consider Figure 1B, where the masking scenario number grows exponentially with respect to the dataset complexity $|X|$ in a combinatorial sense.

Conversely, we raise two related issues to address in the future: (i) designing principles to balance average performance and adaptability to worst-case scenarios (ii) automatically task distribution design in a dataset or instance-wise sense.

Geometric priors can be powerful inductive biases in boosting GenAI

Generally, we refer to constraints or incorporated knowledge in hypotheses space as inductive bias. As stated in Max Welling's comment¹ on the bitter lesson,² machine learning cannot generalize well without inductive biases.

Here, we concentrate on the geometric inductive bias. At a high level, these structures resort to symmetry and scale separation principles,³ particularly necessary in generative modeling of special datasets. For example: human cognitive systems can naturally capture the rotation, translation, reflection, and scaling of signals, implying that the reasonable abstraction of concepts is equivariant to these transformations, as shown in Figure 1C. Another way to apply geometric priors is selective data augmentation by imposing transformation in the data space, meaning that the data themselves can be inductive biases in modeling. Recent advances have verified their effectiveness for scientific discoveries or AI4Science, e.g., molecule design and drug development, better capturing the complex interactions between atoms and predicting the drug candidates.

While geometric priors show promise in GenAI, important questions still need answers to facilitate their use: (i) universal routines to selecting geometric priors in GenAI (ii) effective computation burden alleviation.

Multi-views are required to evaluate performance of models for GenAI

Primarily, GenAI counts more on the data generation mechanism. Given the inherent subjectivity and variability of specific applications, there exist no universally applicable criteria to evaluate generative performance.

For reliability and usefulness, we propose a multi-view evaluation that considers fidelity, diversity, and safety in Figure 1D. The generation fidelity is critical in risk-sensitive applications like dialogue systems in medical science, and standard metrics are log likelihoods or statistical divergence. Diversity is the nature in generation, capturing the complete possible samples. At least two factors influence diversity: observability extent and semantic complexity of the dataset. Observability extent refers to the accessible context information. For example, in image inpainting, the diversity of the generated images decreases as more pixels are observed. Empirically, in large language models (LLMs), increasing the corpus's size and semantic complexity brings more varied and creative text gener-

ation. In Bayesian framework, efficient probabilistic programming requires avoiding posterior or conditional prior collapse to guarantee the diversity.⁴ Additionally, security is an increasingly important concern in GenAI. For example, dataset bias, such as deliberate manipulation or unintentional sampling bias, can significantly affect the LLMs' orientation. Notably, the trend of GenAI is to allow interactions with open environments and continually evolve; securing generative models from attacks at a data level seems urgent. Undoubtedly, there is a solid allure for exploring a model-agnostic and domain-agnostic evaluation schema and integrating multi-views at both the sample and distribution levels.

LS-GenAI IS ON THE WAY

The concept of GenAI has been developed for decades. Until recently, it has impressed us in natural language processing and computer vision, actively engaging in industrial scenarios. Noticing generalization challenges, e.g., limited resources and over-dependencies on scientific discovery empiricism, we propose to scale large models to more practical scenarios with LS-GenAI.

The roadmap of LS-GenAI in Figure 1 relies on the above practical considerations and can be framed as a doubly generative paradigm for simulation and decision-making. Specifically, the identifiable simulation systems or scenarios need to be generated with a few observations (component #1), and the decision-making modules can afford fast adaptation utility in time-sensitive scenarios (component #2), e.g., autonomous driving. At the intersection of simulation science and AI, LS-GenAI has particular use in robotics and life systems, reducing realistic sampling complexity, accelerating scientific progress, and catalyzing discoveries. One prime example of LS-GenAI's potential originates from clinical research. In this context, a high-fidelity biomedical simulation system, operating at the individual level, can create environments to allow the examination of the treatment effects on patients and reduce dependencies on expert experience.

Despite numerous realistic benefits, developing LS-GenAI is nontrivial. The demands of massive real-world data, the lack of high-fidelity world models, and the weak adaptability of these world models have complicated the process of constructing ubiquitous decision-making systems, e.g., in interventional clinical research. In service of the utilities in LS-GenAI, more sophisticated simulation and learning tools must be integrated. Apart from building high-fidelity simulation environments, or world models,⁵ it is essential to support customization for different decision-making tasks such that the task of our interest can be among them. The world exhibits a hierarchical structure, such as the atomic, cellular, tissue, and organismal levels in human body systems or spatiotemporal scales, and retaining multi-scale in generation augmented by symbolic computation can reveal more accurate complex dynamics. Other demands lie in handling partial observability with the inaccessible inherent system state and unpacking the black box to separate function approximation and causal effects. The primary goals of LS-GenAI are to assist in meaningful experimental design and enable fast adaptation of learned skills. Achieving these will ultimately enrich the utilities of GenAI in a broader range of real-world scenarios.

REFERENCES

- Welling, M. (2019). Do we still need models or just more data and compute?. <https://staff.fnwi.uva.nl/m.welling/wp-content/uploads/Model-versus-Data-AI-1.pdf>.
- Sutton, R. (2019). The Bitter Lesson (Incomplete Ideas (blog)), p. 13.
- Bronstein, M.M., Bruna, J., Cohen, T., et al. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. Preprint at arXiv.
- Wang, Y., Blei, D., and Cunningham, J.P. (2021). Posterior collapse and latent variable non-identifiability. *Adv. Neural Inf. Process. Syst.* **34**, 5443–5455.
- Matsuo, Y., LeCun, Y., Sahani, M., et al. (2022). Deep learning, reinforcement learning, and world models. *Neural Network.* **152**, 267–275.

ACKNOWLEDGMENTS

This work is funded by NSFC 62306326 and dedicated to NUDT's 70th anniversary.

DECLARATION OF INTERESTS

The authors declare no competing interests.