Journal Pre-proof

Large-scale generative simulation artificial intelligence: The next hotspot

Qi Wang, Yanghe Feng, Jincai Huang, Yiqin Lv, Zheng Xie, Xiaoshan Gao

PII: S2666-6758(23)00144-3

DOI: https://doi.org/10.1016/j.xinn.2023.100516

Reference: XINN 100516

To appear in: *The Innovation*

Received Date: 22 July 2023

Revised Date: 13 September 2023

Accepted Date: 13 September 2023

Please cite this article as: Wang, Q., Feng, Y., Huang, J., Lv, Y., Xie, Z., Gao, X., Large-scale generative simulation artificial intelligence: The next hotspot, *The Innovation* (2023), doi: https://doi.org/10.1016/j.xinn.2023.100516.

Large-scale generative simulation artificial intelligence: The next hotspot

Qi Wang,[1] Yanghe Feng,[3] Jincai Huang,[3,*] Yiqin Lv,[1] Zheng Xie,[2] and Xiaoshan Gao[4,*]

[1] Kaiyuan Mathematical Sciences Institute, Changsha 410000, China

[2] College of Sciences, National University of Defense Technology, Changsha 410073, China

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

[4] Academy 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.)

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 contents, rather than simply performing discriminative learning tasks.

GenAI's emergence, e.g., the large model, has changed the landscape of deep learning research, 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 artificial intelligence (LS-GenAI) as the next hotspot for GenAI to connect. Throughout the commentary, we use $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$ to denote the explanatory variable, response variable, and latent variable, respectively. For tasks or operators on datasets τ, we represent them in distributions $p(\tau)$.

GenAI can do more than AIGC

Widespread popularity of GenAI models stems from their ability of artificial intelligence 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** data compression, representation disentanglement, and causal inference.

Minimizing the number of required bits to store and transmit information is crucial, known as data compression. 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 or deep variational information bottleneck models, 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 high-dimensional variables and discovering structures of causal graphs for understanding causal effects. Importantly, GenAI with causality enables counterfactual predictions, which renders the potential consequences of a specific intervention that we have yet to execute. For example, with $p\big(y|x,\mathrm{do}(z = z_0)\big)$, policy makers can evaluate the influence of socio-economic policies denoted by z_0, without incurring additional costs.

Despite these fascinating utilities, there remain several tricky questions in the field. (i) Is fully representation disentanglement achievable with generative models? (ii) How

can we identify causal generative models in the presence of small-scale datasets and many unobserved confounders?

Experimental design matters in GenAI's adaptability and robustness

Let us rethink the critical factors contributing to GPT-like models' success. In addition to prompt engineering, the languages' 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 by incorporating new input-output pairs. 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, as aligned with traditional statistical learning theory. However, increasing task complexity requires larger model sizes and comes at the cost of higher computational expenses. For instance, GPT-3 has 175 billion parameters and has been trained on over 570GB of text data from numerous tasks. Generating tasks is problem-specific, and masked learning has emerged as one of the most popular heuristics. 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 $|\mathcal{X}|$ in a combinatorial sense.

Conversely, we raise two related issues to address in the future: (i) Is there a principle to balance average performance and adaptability to worst-case scenarios, particularly when loss values exhibit heavy tails? (ii) How can we automatically design task distributions in a dataset or instance-wise sense to improve generalization?

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 comment[1] on the Bitter Lesson,[2] machine learning cannot generalize well without inductive biases. Inductive biases are particularly beneficial when dealing with data insufficiency, as it guides the learning process in a more reasonable direction.

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. Take the equivariance in symmetry as an 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 data itself can be inductive bias in modeling. Recent advances have verified the effectiveness of geometric priors for scientific discoveries, e.g., molecule design and drug development, better capturing the complex interactions between atoms and predicting the properties of drug candidates. This constitutes a promising avenue for the application of geometric priors in the field of AI4Science.

While geometric priors show promise in GenAI, important questions still need answers to facilitate their use: (i) Are there universal routines to automatically generate geometric priors in GenAI applications? (ii) How can we alleviate the computation burden from constraints or data augmentation?

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 system 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 such as inception score. Diversity is nature of generative modeling, with the purpose of 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 generated images decreases as more pixels are observed. Empirically in large language models, increasing the corpus' size and semantic complexity brings more varied and creative text generation. When using the Bayesian framework, efficient probabilistic programming requires stochastic optimization algorithms to avoid 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 performance and orientation of large language models. Notably, the trend of GenAI is to allow interactions with open environments, incrementally access Internet information, and evolve in a continual learning way; 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 that is end-to-end and integrates 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 with substantial breakthroughs in natural language processing and computer vision, actively engaging in industrial scenarios. Noticing generalization challenges, e.g., limited learning resources and overly 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 above practical considerations and can be framed as the 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 artificial intelligence, 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.

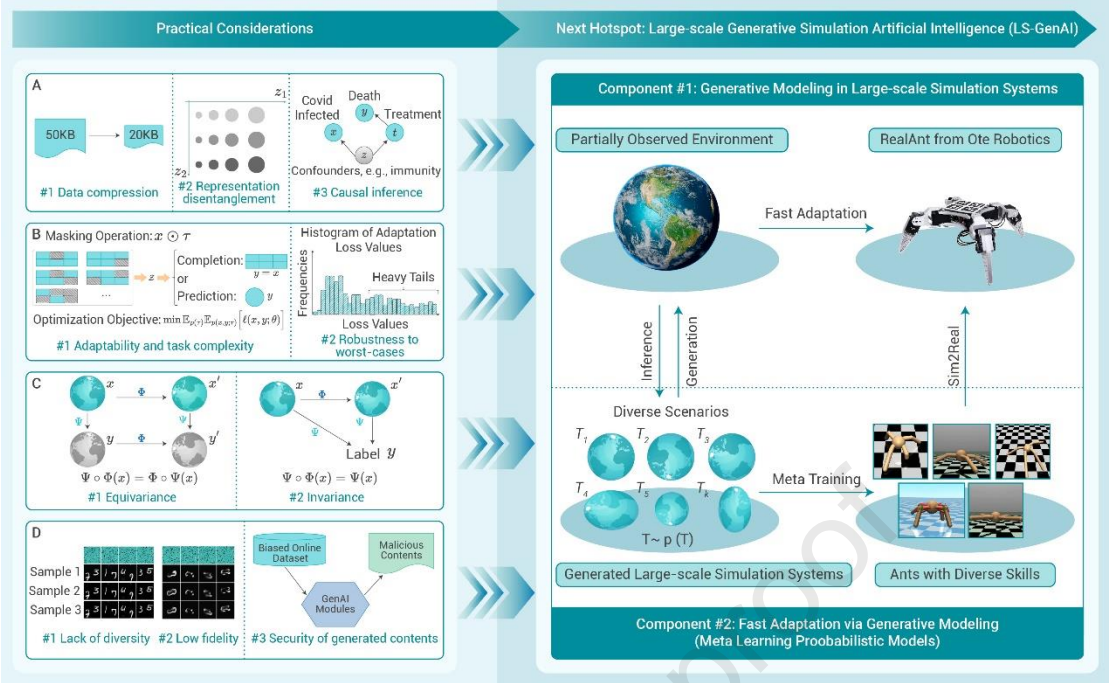

Figure 1. Practical Considerations and the Roadmap of LS-GenAI. The left supports the utilities of LS-GenAI, while the right is the roadmap to achieve via the doubly generative paradigm.

REFERENCES

1. 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.

2. Sutton, R. (2019). The bitter lesson. Incomplete Ideas (blog). 13.

3. Bronstein, M.M., Bruna, J., Cohen, T., and Veli˘ckovi´c, P. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478.

4. Wang, Y., Blei, D., and Cunningham, J. P. (2021). Posterior collapse and latent variable non-identifiability. Advances in Neural Information Processing Systems, ***34***, 5443-5455.

5. Matsuo, Y., LeCun, Y., Sahani, M., et al. (2022). Deep learning, reinforcement learning, and world models. Neural Networks. ***152***, 267–277.

ACKNOWLEDGMENTS

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

DECLARATION OF INTERESTS

The authors declare no competing interests.